Submission to Ecological Psychology. Initially submitted March 23. 2010 (v0.5). Revision v3.4 July 4, 2011

Vocal imitations and the identification of sound events

Guillaume Lemaitre, Arnaud Dessein, Patrick Susini
STMS-Ircam-CNRS, Paris, France

Karine Aura
Laboratoire Octogone-J.Lordat EA 4156, Université de Toulouse, France

Abstract

It is commonly observed that a speaker vocally imitates a sound that she or he intends to communicate to an interlocutor. We report on an experiment that examined the assumption that vocal imitations can effectively communicate a referent sound, and that they do so by conveying the features necessary for the identification of the referent sound event. Subjects were required to sort a set of vocal imitations of everyday sounds. The resulting clusters corresponded in most of the cases to the categories of the referent sound events, indicating that the imitations enabled the listeners to recover what was imitated. Furthermore, a binary decision tree analysis showed that a few characteristic acoustic features predicted the clusters. These features also predicted the classification of the referent sounds, but did not generalize to the categorization of other sounds. This showed that, for the speaker, vocally imitating a sound consists of conveying the acoustic features important for recognition, within the constraints of human vocal production. As such vocal imitations prove to be a phenomenon potentially useful to study sound identification.

Studying vocal imitations to understand sound event identification

Sounds inform listeners about their environment, especially of the part of the environment not currently in their visual field. Think for instance, how annoying it would be to wait right next to a kettle to visually check when the water is boiling. But what exactly do environmental sounds tell us? And what acoustic information do we use to identify

This work was founded by the FP6-NEST-PATH European project n. 29085 CLOSED (Closing the loop of sound evaluation and design). This work was also supported in part by NSF award #0946550 to Laurie Heller.

Portions of this work were presented in July 2008 during the Acoustics 08 Conference held in Paris, France, in July 2009 during the Sound and Music Computing Conference held in Porto, Portugal, and November 2009 during the Auditory Perception, Cognition and Action Meeting in Boston, MA. The authors would like to thank Laurie Heller, Jennifer Tomlison, and Ashlie Henery for proofreading the manuscript.

these sounds? These are two important questions that research in environmental sound perception tries to elucidate. A variety of methods have been developed over the years to question listeners, directly or indirectly, about the acoustic information they use to identify a sound: rating scales, discrimination and categorization tasks, analyses of verbalizations, etc. One aspect of human communication, however, has received little attention: human speakers very often vocally imitate a sound when they want to communicate it to an interlocutor. Although, it is likely that not every sound is vocalizable. But if one assumes that at least some sounds can be vocally communicated, the successful vocal imitation of a referent sound has to convey the acoustic information that is necessary for its identification. In most cases however, the human voice cannot exactly reproduce every feature of a sound. Therefore, one may assume that a speaker will only select the features that he or she deems necessary and sufficient for identification and that he or she can reasonably vocalize well. If these assumptions are correct, successful vocal imitations of environmental sounds (i.e. imitations allowing successful identification) can be considered as a magnifying glass isolating and emphasizing the acoustic information used by listeners to identify sounds. This article examines the potential of using this magnifying glass to cast a light upon the acoustic features that allows sound identification.

Identifying environmental sounds

Empirical studies have consistently shown that listeners can identify what has caused a sound. Whereas some types of information about the sound source are conveyed by the sound only, others can not be interpreted without additional knowledge provided by the context.

A sound is an audible acoustic wave caused by a mechanical sound event: the air blowing out of a kettle, a pencil falling from a desktop, the membrane of loudspeaker vibrating. Now, consider the following descriptions of the same sound provided by different listeners (Lemaitre, Houix, Misdariis, & Susini, 2010): "a series of short high-pitched sounds", "a few small objects dropped onto a glass or ceramic surface and bouncing", "a couple of ice cubes dropped into glass". They all are correct, and illustrate that listeners are capable of both analyzing the properties of the acoustic wave (e.g. pitch, timbre, temporal sequencing), recovering the basic mechanical interactions (e.g. smalls objects dropped onto something hard), and interpreting the *source* of the sound. But this latter interpretation requires other available sources of knowledge (Gaver, 1993a, 1993b; Handel, 1995). For instance, if the listener is aware that the situation takes place in a kitchen, he or she will likely interpret the small objects as ice cubes poured into a glass. But listeners not aware of the context may have a different interpretation: "coins dropped on a ceramic dish", "aspirin tablets dropped into a cup", etc. (these were descriptions from the same study). Some properties are unambiguously specified by the sound only (e.g. dropping, small objects, a resonant surface), but some others are interpreted based on the context: were the small objects ice cubes, aspirin tablets, coins? Were they dropped onto a flat surface or into some sort of container? Was it made of glass, metal, or ceramic?

These questions are found throughout the literature on environmental sound perception. Vanderveer (1979) was the first to show that listeners spontaneously describe the cause of the sounds they hear. Many authors have shown the listeners' ability to auditorily recover the properties of isolated simple sound events: size and shape (Lakatos, McAdams,

& Caussé, 1997; Carello, Anderson, & Kunkler-Peck, 1998; Kunkler-Peck & Turvey, 2000; Houben, Kohlrausch, & Hermes, 2004; Grassi, 2005), material (Klatzky, Pai, & Krotkov, 2000; McAdams, Chaigne, & Roussarie, 2004; Giordano & McAdams, 2006; McAdams, Roussarie, Chaigne, & Giordano, 2010), action (Warren & Verbrugge, 1984; Cabe & Pittenger, 2000).

In many cases however, it has been very difficult to pinpoint the acoustic features that specify these properties to the listeners. A typical example is that of the perception of the material of impacted objects. There exists in theory an acoustic property that specifies unequivocally the material of such an object: the ratio of each partial's decay time over its frequency (Wildes & Richards, 1988). In the sense of the ecological approach to perception, it is an invariant (Carello, Wagman, & Turvey, 2005). However, empirical studies have shown that listeners do not use this property to judge the material, and that they have only a coarse ability to distinguish the material of the sounding objects when other properties are varied (see Klatzky et al., 2000; Lutfi & Oh, 1997; Giordano, McAdams, & Rocchesso, 2010; Lemaitre & Heller, 2011).

As it turns out, interpreting meaningful sounds (sounds that listeners can identify the cause of) shares a lot with the processing of language (Howard and Ballas, 1980; Ballas and Mullins, 1991; Ballas, 1993; Cummings et al., 2006). Ballas and Howard (1987) showed for instance the importance of the phenomenon of homonym-like sounds: sounds that can be discriminated (subjects can tell that the sounds are different), but identified as the same event (the cause appears to be identical). The sounds of a fuse burning and food frying are examples of two homonymous sounds.

Many pieces of information that a listener can interpret from listening to a sound therefore are inferred from more than just the sound. Before trying to highlight the acoustic features used by listeners to identify a sound source, it is first important to consider which perceived properties of the sound events can possibly be conveyed by acoustic features only. Our previous work has shown that the perceived sound events are organized in listeners following a taxonomy that is very close to that proposed by Gaver (1993b), with a first separation between actions made by solid objects, liquids and gas, and a special emphasis on the temporal aspects of the elementary actions (e.g discrete vs. continuous actions, etc. Houix, Lemaitre, Misdariis, Susini, and Urdapilleta, 2011). Furthermore, we have also shown that listening to environmental sounds activates in listeners lexical representations of the elementary mechanical actions that have caused the sounds (e.g. tapping, scraping, rolling, blowing, dripping, etc.) more strongly and rapidly than the many other possible descriptions of the sources of the sound, even without the help of the context (Lemaitre & Heller, 2010). Therefore, the present study used a set of sounds that were easily identifiable at the level of the elementary actions organized in this taxonomy (Lemaitre et al., 2010).

Vocal imitations as a method to assess sound perception

Different methods are commonly used to assess which acoustic features listeners use to identify the sounds. A widespread psychoacoustic technique consists in synthesizing sounds and varying the parameters expected to subserve identification. But such a method is not available when the experimenter wants to use recordings of natural events. In such cases, another method consists of first using dissimilarity ratings and multidimensional scaling analysis to identify the relevant features, and second typicality judgments to map the

features to the different categories (Lemaitre, Susini, Winsberg, Letinturier, & McAdams, 2007, 2009). Such a technique is however only appropriate for sounds caused by similar sources (Susini, McAdams, & Winsberg, 1999). For heterogeneous sets of sounds, free categorization and linguistic analyses of the listeners' descriptions of their categories are generally used (Houix et al., 2011). Among the various linguistic devices used to describe a sound, we have observed that vocal imitations are spontaneously used when subjects have to communicate a sound that they have just heard (Lemaitre, Dessein, Aura, & Susini, 2009), as other authors have also noticed (Wright, 1971). Specifically, when no proper vocabulary is available, vocal imitations may facilitate the communication of an acoustic experience. This is exactly what people do when they call the "Car Talk" radio show, and try to vocalize the sound that their car is making to presumably indicate a problem to the hosts[1].

Vocal imitations, onomatopoeias and sound symbolism

In fact, there are two different types of imitations: imitations standardized in a language (onomatopoeias) and non-conventional and creative vocalizations. Imitations of the former type are close to words: the meaning is associated to the word through a symbolic relationship (Hashimoto et al., 2006). Our study is more interested in the latter kind, for which the meaning is conveyed by some similarity between the imitation and what it imitates. Let us first examine the difference between these two types of imitations.

Onomatopoeias have probably been the most commonly studied type of vocal imitations. Pharies (1979, cited by Sobkowiak, 1990) provided a very interesting definition:

> An onomatopoeia is a word that is *considered by convention* to be *acousticly similar* to the sound, or the sound produced by the *thing* to which it refers.

The sound symbolism of onomatopoeias has been studied for several languages (Sobkowiak, 1990; Rhodes, 1994; Oswalt, 1994; Żuchowski, 1998; Patel & Iversen, 2003). In particular, Japanese onomatopoeias have been much studied. For instance, Iwasaki, Vinson, and Vigliocco (2007) have experimentally shown that English listeners (with no proficiency in the Japanese language) would correctly rate the meaning of *giongo*: common onomatopoeias mimicking sounds. Systematic relationships between the aesthetic impressions, phonetical content and acoustic properties have been highlighted for Japanese onomatopoeias (Takada, Tanaka, & Iwamiya, 2006; Takada, Fujisawa, Obata, & Iwamiya, 2010).

In comparison to onomatopoeias, non-conventional vocal imitations have been rarely studied. Such imitations can be simply defined by dropping the first part of Pharies' definition of onomatopoeias: a non-conventional imitation is a creative utterance intended to be acoustically similar to the sound, or the sound produced by the thing to which it refers. Therefore, a non-conventional imitation is only constrained by the vocal ability of the speakers, and does not use symbolic conventions. Lass et al. (1983) showed that human-imitated animal sounds were well recognized by listeners, even better than the actual animal

[1] `http://www.cartalk.com/`. For instance, in a recent show:
"- So, when you start it up, what kind of noises does it make?
- It just rattles around for about a minute. Just like it's bouncing off something. He thinks that it could be bouncing off the fan, but it's not there. [...]
- Just like bududblu-bududblu-bududblu?
- Yeah! It's definitively bouncing off something, and then it stops."

sounds (Lass, Eastham, Parrish, Sherbick, & Ralph, 1982), yet the listeners did not have any problem discriminating between the two categories (Lass et al., 1984). This effect is probably close to that of the Foley sound effects used in movies and video games (Heller & Wolf, 2002; Newman, 2004). The meaning of "tame" sound symbolisms (i.e. onomatopoeias) may be specific to a culture, whereas "wild" sound symbolisms (non-conventional) are imitative rather than symbolic, to borrow the words of Rhodes (1994). Therefore, our study focused on subjects using only wild imitations.

Vocal imitations have been also used to develop technical applications (Ishihara, Tsubota, & Okuno, 2003; Ishihara, Nakatani, Ogata, & Okuno, 2004; Nakano, Ogata, Goto, & Hiraga, 2004; Nakano & Goto, 2009; Sundaram & Narayanan, 2006, 2008; Takada et al., 2001; Gillet & Richard, 2005). For instance, using vocal imitation as a control of sound synthesis is a promising approach (Ekman & Rinott, 2010). *Cartoonification* is another specific kind of sound synthesis that consists of exaggerating some acoustic features (perceptually important) while discarding some (Rocchesso, Bresin, & Fernström, 2003).

But is any kind of sound vocalizable? Besides speech, humans can produce a wide variety of vocal sounds, from babbles to opera singing, from sighs to yells, from laughter to gurgles. Beatboxers[2] have developed vocal techniques that allow them to imitate the sounds of drums, turntables and other sound effects commonly found in popular music, with a proficiency that compares only to the lyrebird (Proctor, Nayak, & Narayanan, 2010). Despite these somewhat extraordinary performances, several limitations are to be considered. First, there are physiological limitations. The voice apparatus can be essentially approximated by a source-filter model, with the lungs and the vocal folds as the source (i.e. the glottal signal), and the articulators (vocal tract, tongue, palate, cheek, lips, teeth) as the filter. The main limitation to what the voice can do probably comes from the glottal signal. The glottal signal is produced by a single vibrational system (the vocal folds), which implies that vocal signals are most often periodic (even though, chaotic, a-periodic or double-periodic oscillations can also happen), and essentially monophonic (even though some singing techniques can produce the illusion of multiple pitches). Furthermore, its pitch range is limited. The range of the human voice extends overall from about 80 Hz to 1100 Hz, and a single individual's vocal range usually covers less than two octaves. Another kind of limitation comes from speakers's native language. Speakers have a better ability to produce the speech sounds of their native language, and usually encounter utter difficulties when attempting to produce the sounds of a foreign language (Troubetzkoy, 1949; Strange & Shafer, 2008). For instance, the French speakers used in this study, even if instructed not to use words, were of course more prone to produce French trills /r/ and /ʀ/[3] than the Spanish trill or the English /ɹ/, and very unlikely to use the English dental fricatives /θ/ and /ð/. A last limitation comes from the fact that some speakers may be better able to invent successful imitations of a sound than some other ones.

Outline of the study

The goal of the research reported here was to answer two questions: can the vocal imitations of sounds made by novice imitators allow a listener to recover the imitated sound?

[2]For a compelling example, see `http://www.neurosonicsaudiomedical.com/`

[3]We did not distinguish between /ʀ/ and /ʁ/

For the sounds that are successfully vocalizable, could we study vocal imitations to better understand what acoustical features are necessary to identify the sounds sources? As a starting point, we used the sounds of categories of simple mechanical events previously mentioned that our previous work showed to be well identified on the basis of their acoustic features, and require little semantic interpretation.

The experiment reported in this paper required a set of participants to vocally imitate these sounds, and then another set of participants to categorize these imitations in terms of what they thought was imitated. We used novice speakers and naive listeners (i.e. participants with no expertise in sound analysis or vocal production), because we were interested in how effectively these vocal imitations could communicate another sound, and not in their realism (the imitations of expert practitioners have sometimes been rated as more realistic than the actual sounds of the events - Heller & Wolf, 2002). The comparison of the two categorizations (referent sounds vs. corresponding imitations) showed that the listeners were able to recover the categories of sound events to which most vocal imitations corresponded. This showed that vocal imitations enabled the identification of many sound events. We then sought to identify the information in the successful vocal imitations that conveyed the referent sound events. The data were submitted to a simple machine learning technique (binary decision tree). Finally, we tested if the classifier, trained on the vocal imitations, could predict the classification of the referent sounds, and other sounds from the same categories as well.

An experimental categorization of vocal imitation

This experiment focused on how listeners categorized a set of vocal imitations of kitchen sounds. More precisely, we studied here only non-linguistic (wild) imitations. Using a categorization task was motivated by the assumption that, if listeners are able to recover the sounds that are imitated (the referent sound), they should categorize the imitations in a similar way to what they would do with the referent sounds. If kitchen sounds are categorized according to the corresponding sound events (water flowing, food being cut, etc.), so should their imitations. Furthermore, because categorizing a set of sounds can be done at different levels of specificity, using a categorization task allowed us to explore what kind of information was conveyed by the imitations: for instance, a listener might not be able to identify the cause of the sound at the most specific level, but still be capable of recovering a more general category which the sound event belongs to.

Initial data: the categorization of kitchen sounds

The stimuli used in the experiment were based on vocal imitations of a set of everyday kitchen sounds. The selection of these sounds was made on the basis of the results of a categorization experiments reported in Houix et al. (2011). Below, we first reproduce the main outlines of this study, and describe the results that were relevant to our study.

Procedure. The initial sounds were 60 recordings of activities usually occurring in a kitchen, chosen from different commercial sound libraries. The participants had to listen to the sounds, group them together, and describe their grouping, following the procedure described in the following paragraph. We used the data collected for 15 non-expert participants, because the results of Lemaitre et al. (2010) showed that only non-expert participants

used a classification strategy consistently based on the identification of the sound events (which is the focus of this study), as opposed to a classification based on acoustic similarities irrespective of the cause of the sounds.

Analysis and sound selection. An analysis of the categories and their descriptions can be found in Houix et al. (2011). The upper panel of Figure 1 represents the results of the categorization.

On this figure, we have represented the nine clusters that were highlighted by the linguistic analysis of the subjects' verbalizations reported in Houix et al. These clusters corresponded to sounds made by cutting, preparing food, impact on pots and pans, crumpling/crushing, closing, machines, water, cooking and gases. Therefore, they each corresponded to a specific kind of mechanical event: single impact, repeated impacts, drips, gusts, etc. In the experiment reported here, we considered the four categories of sounds of liquids (water), gases, electrical appliances (machines), and solids (cutting). The upper panel of Figure 1 shows that these sounds were consistently clustered. We chose three sounds in each of these categories (see Table 1). These sounds were in general well identifiable[4], although for the electrical sounds we had to choose sounds that were slightly less identifiable, because of the unavailability of more identifiable sounds in this category. The lower panel of Figure 1 represents the hierarchical tree obtained for the 12 sounds used as referent sounds only, using the same type of cluster analysis as we used to analyze the classification of the imitations (see below). This figure shows that these 12 sounds formed 4 clearly distinguishable categories. These sounds could therefore be described at two levels of specificity: the type of sound production (solid, electric, liquid, gas), and the specific source (cutting food, gas stove, etc.).

Imitating the kitchen sounds

Method. To provide the material for the subsequent experiment, 20 participants were hired (10 men and 10 women, aged from 18 to 50 years old). The participants were seated in a sound attenuated booth, and required to listen to the sounds. For each sound, they had to record three instances of an imitation of the sound. They were required to imitate the sounds "in such a way that another person could recognize it". They were instructed not to use any words or onomatopoeias. They were alone in the booth, using a specifically designed Max/MSP interface. They could listen to their imitations and discard those they did not like. The sounds were recorded using a Schoeps MK5 microphone and a RME Fireface 400 sound board. The sounds were recorded at 44.1 kHz/16 bits resolution.

A total of 720 (4 categories x 3 sounds x 20 participants x 3 trials) imitations were recorded. These recordings were edited and screened to remove those that were of poor quality, and those that included words or onomatopoeias. Only the best of the three trials was selected for each participant and each sound (as participants were recording themselves, many recordings were not of sufficient quality, due to being too close to or too far from the microphone or stopping the recording in the middle of the imitation, etc.). Eventually, only

[4]The subjects' confidence in identifying the cause of the sound was measured (see Table 1); the mean value was 6.07 on a scale ranging from 0 - the participant does not know at all what has caused the sound - to 8 - the participant perfectly identifies the cause of the sound.

the imitations of six participants (three men and three women) were selected, making a total of 72 imitations.

The referent sounds and their imitations. The referent sounds are described in Table 1, and their spectrograms are represented in Figure 2. Within the broader categories of types of interaction, the sounds shared some apparent similarities. The three sound of electrical appliances were produced by the rotation of a motor and the occasional the impact of bits of food hitting an object such as a bowl. These sound therefore consisted of a low steady fundamental frequency and occasional transients. The three sounds of gas had their noisy spectrum in common.. The three sounds of solids all consisted of the repetition of a very brief impact-like element. The similarities between the three sounds of liquid were less evident: for instance, the steady broadband noise resulting form water gushing from a tap was acoustically very different from the three isolated chirps caused by water dripping in a container.

The 72 imitations are provided as supplemental material. Table 2 reports a phonetic transcription of these imitations. These transcriptions are approximate (and sometimes impossible), because in many cases the imitators used other sounds than the phonemes of French speech.

These transcriptions, however, showed some interesting properties. First, the imitations were mostly made of voiceless consonants (/ʃ/, /f/, /k/, /p/, /s/, /t/) and the uvular and front trills (/ʀ/ and /r/) common in French. Second, the imitations of some sounds shared apparent phonetic similarities. For instance, most of the imitations of the sound E2 consisted of the uvular trill /ʀ/, repeated or prolonged by an elongated vowel. It is also important to note that these sounds imitated the sound of a food processor, and that the imitators were rather successful at capturing the pitch of this sound (an aspect not apparent in the phonetic transcription). Another interesting example is the sound L2. It consisted of three drips, and the six imitations were all made of three repetitions of a similar element (often beginning with the voiceless bilabial stop /p/). More generally, it appeared that the imitations of a same specific sound or a same category of sound production sounded somewhat similar, and that the temporal aspect of the imitations (pitch contour, repetitions, etc.) was an important component of the similarities between the imitations of a similar (or identical) sound. These temporal aspects seemed to capture the main similarities between the sounds of category described in the previous paragraph. The experiment reported in the next paragraph investigated whether a set of listeners were able to use these similarities to recover the referent sounds.

Free classification of imitations: method

Participants Twenty participants (10 women and 10 men) volunteered for the experiment and were compensated for their participation. Ages ranged from 18 to 50 years old. All reported having normal hearing and were native French speakers. They had no previous experience in vocalization.

Stimuli The 72 vocal imitations described in the previous paragraph were used as stimuli.

Apparatus The sounds were played on a Macintosh Mac Pro (Mac OS X v10.4 Tiger) workstation with a MOTU firewire 828 sound card. The stimuli were amplified diotically over a pair of YAMAHA MSP5 loudspeakers. Participants were seated in a double-walled IAC sound-isolation booth. The study was run using Matlab.

Procedure The participants were all given written instructions (in French) explaining the sorting task. They saw a white screen, on which red dots labelled from 1 to 72 were drawn, each dot corresponding to a sound. The labeling was different for each participant. They could hear the sound by double-clicking on a dot. Participants were asked to move the dots to group the sounds together. They were allowed to form as many groups as they wished and to put as many sounds in each group as they desired. Participants were required to group together the vocal imitations "on the basis of what is imitated". Specifically, they were warned not to categorize the speakers. After they had made the categories, they had to describe the categories to the experimenter.

Free classification of imitations: analysis

Descriptions of the categories. Although the descriptions provided by the participants were collected using an informal method, and could therefore not be systematically analyzed, they nevertheless provided us with some useful indications regarding the strategies used by the participants. Using the typology established by Lemaitre et al. (2010), these descriptions suggested that the participants used different kinds of similarities to group sounds together. Most of the verbalizations described *causal* and *semantic* similarities (i.e. the causal event, the interpreted source and the meaning associated with it). But other kinds of similarity were also used: acoustic properties of the sounds, feelings (called here *hedonic* properties), and, more rarely, similarities in the vocal production of the imitations (see Table 3 for an example of such descriptions). For some participants, the description of a given class mentioned several kinds of similarity (e.g. "Continuous sounds, with a kind of vibration, with the lips, the throat, there is something spinning, noises of machines").

Individual differences and potential outliers. These descriptions might therefore indicate different strategies across the participants. In particular, the descriptions of one participant appeared rather incoherent. Appendix A reports the method that we used to identify individual differences and outliers. As a result, we eliminated one participant from the analyses.

Confusion matrix. A first overview of the participants' classification is provided by the confusion matrix represented in Figure 3. In this figure, each cell of the matrix corresponds to a pair of sounds, and its color represents the number of participants who placed these two sounds in the same group. If the six imitations of each referent sound would be systematically and exclusively grouped together (i.e. with no other sounds than imitations of the same referent sound) only the 6x6 submatrices along the main diagonal would be dark. If the imitations were only grouped together with imitations of referent sounds from the same category, only the 18x18 submatrices along the main diagonal would have been dark. Instead, the pattern of the confusion matrix highlights different phenomena for the four initial categories of referent sounds. The imitations of each of the three sounds of

gases were rather consistently grouped together, or with imitations of other gases. A few errors resulted from the improper grouping with imitations of liquids. The imitations of the sounds of electrical appliances and of the solids were not as systematically grouped together. Rather, all the classifications resulted from grouping the imitations with imitations from the same categories, but not necessarily with an imitation of the exact same sound. A few misclassifications with imitations of liquids also occurred. Finally, with the exception of the imitations of the sound L2, the imitations of liquids were all misclassified.

Analysis of the classification. These initial insights were confirmed by submitting the data to a hierarchical clustering analysis, and representing the similarities between the sounds (two sounds are similar when they have been grouped together by a large number of participants) in a dendrogram.

To identify significant clusters in a dendrogram, the dendrogram is usually cut at a given fusion level. As an alternative clustering method, we used a threshold of *inconsistency*. The advantage of using the inconsistency coefficient is that it emphasizes compact subclasses that would not be revealed using the fusion level (see Appendix B for a description).

The dendrogram of vocal imitations is represented in Figure 4 (using an unweighted average linkage method). The coefficient of cophenetic correlation is 0.95, indicating that the dendrogram fairly represents the proximity data. The indexes on the x-axis correspond to the vocal imitations (the first letter corresponding to the speaker, the second to the imitated sound). The branches in the gray rectangles correspond to the different clusters highlighted by setting the threshold of inconsistency to 1.45. The threshold was set by decreasing the inconsistency, and so increasing the number of clusters, until having a set of clusters of imitations interpretable in terms of categories of referent sounds. The rectangles are indexed by script letters. When a cluster includes only imitations of a same referent sound, the index of clusters also receives the number corresponding to the referent sound.

Exploring the clusters of vocal imitations highlights the principles that ruled the categorization of the vocal imitations, as well as the characteristics of the imitations potentially responsible for the clustering. Considering the dendrogram from the highest fusion level, the first division (A) distinguished the imitations of gases from all the other sounds. The former imitations were clearly distinct from the others because of their breathy (unvoiced) character. The latter imitations were further divided into two clusters (division B): on the left hand side, a cluster that includes a subcluster (division C) mostly consisting of electrical sounds (characterized by the presence of a continuous steady pitch), and a hybrid subcluster that mostly includes imitations of liquid sounds (sound with a rhythmic pitch). On the right hand side a cluster was further subdivided (division D) into a subcluster of imitations of solid sounds, and a hybrid cluster of liquid and solid sounds. These imitations all had a repetitive pattern in common. Thus, the division of the dendogram results in four distinct and coherent clusters: the imitations of gases, electrical sounds, of some liquid sounds, and sounds of solid objects. The other imitations of liquids are rejected in a hybrid cluster, or mixed either with imitations of solids, or imitations of electrical sounds.

At the finest level of the hierarchy highlighted by the threshold of inconsistency, eight clusters are both mathematically consistent and consistently related to the referent sounds. These eight clusters[5] correspond to the four main categories of referent sounds:

[5]When a cluster only incorporates imitations of the same referent sounds (e.g. G_1), this cluster receives

- $\mathcal{G}$: imitations of gases
  (1) $\mathcal{G}_1$ made of the 6 imitations of the sound G_1;
  (2) $\mathcal{G}_2$ made of 5 of the 6 imitations of the sound G_2;
  (3) $\mathcal{G}_3$ made of 5 of the 6 imitations of the sound G_3;
- $\mathcal{E}$: imitations of electrical sounds
  (4) $\mathcal{E}$ made of 12 of the 18 imitations of the sounds E_1 (4 imitations), E_2 (2 imitations) and E_3 (6 imitations);
- $\mathcal{L}$: imitations of sounds of liquids
  (5) $\mathcal{L}'$ made of 6 of the 12 imitations of the sounds L_1 (4 imitations) and L_3 (2 imitations);
  (6) $\mathcal{L}_2$ made of 6 of the 6 imitations of the sound L_2;
- $\mathcal{S}$: imitations of sounds of solid objects
  (7) $\mathcal{S}''$ made of 8 of the 18 imitations of sounds S_1 (2 imitations), S_2 (3 imitations) and S_3 (3 imtations);
  (8) $\mathcal{S}'$ made of 5 of the 12 imitations of sounds S_1 (3 imitations) and S_2 (2 imitations).

One cluster ($\mathcal{X}$) is mathematically consistent, but includes imitations of different sounds.

For most of the referent sounds, the majority of imitations were categorized in easily interpretable clusters (i.e. clusters including only imitations of the same referent sound, or grouping imitations of sounds from the same original category). The clusters $\mathcal{S}'$ and $\mathcal{S}''$ mix imitations of the three referent solid sounds. Two sounds lead to imitations inappropriately clustered: the sound E2 (food processor) lead to imitations either clustered with the imitations of the sounds E1/E3, L1/L2/L3, or were not clustered. The sound L3 (water running in a sink) lead to many imitations that were not consistently clustered. If it is likely that these sounds were probably more difficult to vocalize, it is also worth noticing that the sound E2 was not as recognizable as the other sounds (see the confidence values in Table 1). In this particular case, a further assumption might be that the participants could not decide upon which feature to emphasize when imitating, because they were not able to recognize the original sound.

Overall, 58 imitations (out of the 72) fall in the four clusters $\mathcal{G}$, $\mathcal{E}$, $\mathcal{L}$ and $\mathcal{S}$ made by grouping together the eight mathematically consistent clusters of imitations, and corresponding to the categories of referent sounds. Among these 58 imitations, only three are clustered in a cluster that does not correspond to the category of referent sounds. Therefore, if we consider the categories of referent sounds as an appropriate level of accuracy, 55 of the vocal imitations (76.4 %) were consistently classified. This indicates that, for a large majority of the imitations, listeners were able to access the category of the referent sound.

Looking at the dendrogram from the top, its superstructure also presents a number of important differences from that in Figure 1. Whereas for the referent sounds the first division of the dendrogram separated the sounds of solids from all the other ones, the first division in Figure 1 distinguishes the imitations of gases. As indicated above, these imitations had a clear distinct breathy character that is not shared by any other sounds. Most of the imitations of liquids were inconstantly categorized.

---

the same name as the referent sounds (e.g. $\mathcal{G}_1$). When a cluster incorporates imitations of referent sounds belonging to a similar categories (e.g. liquids), the cluster received the initial of this category (e.g. $\mathcal{L}'$).

Discussion

The results of the experimental categorization of the imitations has shown that the listeners were able to categorize the imitations in clusters that were consistent with the categories of referent sounds, with the exception of the imitations of the liquids. These categories grouped together sounds caused by similar mechanical events. Therefore, this indicates that the participants were able to recover the categories of referent sound events, for most of the imitations. They could just as well have not recovered anything from the referent sounds, or, despite the instructions, chosen to make categories on the basis of only local similarities between the vocal sounds, not relevant to the referent sound events, if such an organization would have been an easier principle to categorize the sounds.

However, this classification was not perfect. Except for the gases, the accuracy of identification was limited to the categories of referent sounds, and not to the referent sounds themselves: the clusters included the imitations of the different referent sounds form the same category. Furthermore, the results suggest a mix of different strategies of classification as is also the case in the categorization of kitchen sounds reported by Lemaitre et al. (2010) and ? (?).

In particular, the imitations of liquids were inappropriately categorized. In the dendrogram of the referent sounds, the liquids were clearly distinct from the other categories. However, whereas the imitations of the sound L2 form a rather stable cluster (though associated with the clusters of imitations of solids), the imitations of the sounds L1 and L3 were aggregated with the imitations of electrical appliances, without creating any stable cluster. The sounds L1 and L3 were long continuous steady sounds (and so were their imitations), and this characteristic might have made them close to the imitations of electrical sounds, which display a continuous steady hum. The sound L2 was made by a series of water drips. The imitations of this sound presented a rhythmic pattern that might have made them close to the sounds of cutting food (i.e. the solids). However, despite these patterns, the referent sounds of solids and liquids were categorized in distinct categories. This suggests that the listeners who categorized the referent sounds were able to use some other cues that were not present in the imitations. Sounds of liquids are in general characterized by the presence of bubbles. These bubbles result in rapid frequency sweeps (chirps), occurring on top of noisier and louder components. One possible interpretation is that the imitators were unable to vocally render both the noisier components and these rapid sweeps, probably because of the physiology of the vocal apparatus. The listeners were therefore not able to pick up the "liquid" identity of these sounds, and instead grouped them with imitations presenting some other irrelevant superficial similarity, conveyed by the time course of the sounds. It is interesting to note that the imitations of the sound L2 (three drips) were all grouped together, and that several subjects used vowel glides that somehow rendered the drips (e.g. [ui] or [wi]).

Examining the structure of the resulting dendrogram in fact suggested that the clusters of imitations of a same category of events overlapped with a set of simple apparent characteristics of the imitations: noisy vs. periodic, continuous vs. rhythmic, etc. To verify if such acoustic characteristics could be the basis for the categorization of the imitations, we report in the next section how we trained an automatic classifier to categorize the imitations in a way similar to what the listeners did. However, even if the experimental categorization

may be predicted on the basis of a few of acoustic features, this does necessarily mean that these features can be considered as characterizing the different categories of sound events, generally. It could also be possible that the participants used an ad hoc strategy in our experiment, and only picked up the features that were the most efficient to form a few of contrasting classes in this *particular* set of sounds. In fact, some authors have suggested that listeners can use any configuration of information that is potentially useful to achieve a particular task (rather than using predetermined features for predetermined tasks – see Handel, 1989). For instance, Aldrich, Hellier, and Edworthy (2009) showed that subjects used different features to achieve different tasks, and that the features they used were even influenced by the particular set of sounds used in the experiment. Along the same idea, McAdams et al. (2010) showed that listeners presented with the same set of sounds reconfigured how much they weighed different acoustic features depending the experimental task (similarity judgment or categorization task), and that different subjects would differently weigh the features for the same task. Giordano, McDonnell, and McAdams (2010) showed that listeners required to do a task in fact used a continuum of information, rather than just picking up only the most relevant features.

To explore the features of the imitations that the listeners used we trained an automatic classifier to select the most relevant features to predict the classification of the imitations, and applied it to the referent sounds.

Predicting the classification from the acoustic features

The results of the categorization task were first used to fit a model that predicts the categories on the basis of some acoustic features of the imitations, thus confirming the subjects in the classification experiment grouped together imitations that shared a few of acoustic features.

Acoustic properties of the clusters of imitations

The description of the clusters of imitations suggested that they might be characterized by a few distinctive acoustic features. To uncover such features, the data were submitted to a binary decision tree analysis. The goal of this analysis was to recursively predict each division in the dendrogram of imitations by a set of binary decision rules based on a few acoustic features. Although much more sophisticated automatic classification techniques are available, and would probably perfectly learn how to automatically classify our set of vocal imitations, binary decision trees have the advantage of being very simple: each distinction between two classes is predicted by the combination of independent binary rules (e.g. the sounds belong to cluster C if feature $F_1 < threshold_1$ and feature $F_2 > threshold_2$). If not the most powerful, the results of a binary decision are nevertheless easily interpretable, and therefore more appropriate to uncover the categorization principles. A binary decision tree is therefore a conveniently simple model to fit the data, if we assume that the distinctions between the most consistent categories of imitations found in our data are clear-cut and depend only on a few acoustic properties of the imitations.

We will describe in the following paragraphs the result of the analysis for each division of the dendrogram (the indexes of the divisions make reference to Figure 4). For each division, we have first considered the imitations that fell in consistent clusters. These

clusters of imitations were submitted to the algorithm that selects the acoustic features that best predict the division. We have then shown the imitations that could not be consistently grouped together with any of the clusters where were located.

In many cases, several acoustic features can do the job of predicting a division equivalently well. We have reported here only the features that were meaningful, i.e. features that could be interpreted in terms of how the sound event might have structured the referent sounds. For instance, we did not report the "blind" statistics of the signal (Mel-frequency spectral coefficients, etc.), even though they could also predict the classification. In fact we were not interested in the computational power of the automatic classifier, but in what it might reveal about the information that the imitators tried to convey to the listeners.

The acoustic features were computed with the IrcamDescriptor toolbox (Peeters, 2004), and the Yin algorithm to compute the fundamental frequency and the aperiodicity of the signal (de Cheveigné & Kawahara, 2001). For this last algorithm, we set the threshold of aperiodicity that allows the computation of a pitch to 0.5 (the usual value is 0.1), because many of our signals included noisy whistles, the pitch of which would have otherwise been missed. To ensure that the resulting fundamental frequencies were actually measuring the frequency of some periodic part of the signal, we manually inspected each sound with Praat (Boersma & Weenink, 2009). These algorithms compute a value of each acoustic features for a number of time frames along the duration of each sound. We have used here only the statistics summarizing these features (average, standard deviation, extrema, etc.).

Division A: imitations of gases vs. all the other ones. Division A separates the imitations of gases from all the other imitations : $\mathcal{G}_1$, $\mathcal{G}_2$, $\mathcal{G}_3$, $\mathcal{X}$ vs. $\mathcal{E}$, $\mathcal{L}_2$, $\mathcal{S}$, $\mathcal{S}'$. We considered all the 72 sounds here. Two features allowed one to discriminate between these two classes (see the upper panel of Figure 5): the average *fundamental frequency* and the *modulation amplitude of the energy envelope*. The fundamental frequency is an acoustic feature that correlates with the sensation of pitch (Marozeau, Cheveigné, McAdams, & Winsberg, 2003). The value considered here was averaged over all the periodic parts of the signal. The *modulation amplitude* measures how steady the energy envelope of a sound is. For instance, all the sounds that displayed the repetition of an element separated by periods of silence had a high modulation amplitude. The figure shows that the imitations of gases were separated from all the other sounds because they were steady signals presenting a high fundamental frequency (a hiss). Seventy-one of the seventy-two imitations were correctly classified (98.6%). The only misclassified sound (M_1L_3) was an imitation that had two different parts: a first part made of a low hum, followed by a breathy hiss. Because we were only using values averaged over the duration of the signals, we were not able to handle such sounds.

Division B: Electrical appliances vs. solids. Division B separates the imitations of electrical appliances (and some liquids) from the imitations of solids (and some other liquids). Fifty sounds were considered here. Two features allowed one to predict this division: the modulation amplitude (already mentioned), and the *effective duration* of the sounds (see the middle panel of Figure 5). Accordingly, the clusters $\mathcal{L}2$, $\mathcal{S}'$, and $\mathcal{S}''$ were separated from the clusters $\mathcal{E}$ and $\mathcal{L}$ because their energy envelope was modulated (remember that the solids were sounds of food being cut whereas the electrical appliances were steady sounds

of motor), and because the sounds of electrical appliances had a longer duration (note that the latter feature is only useful for three sounds). Forty-nine of the fifty imitations were correctly classified (98%). The only sound that was misclassified was an imitation of a solid (M_3S_1) that was in fact misclassified in the dendrogram of imitations (it did not belong to any stable cluster).

Division C: Electrical appliances vs. liquids. Division C separates the imitations of electrical appliances from the imitations of liquids. If we consider only the 20 imitations that fell in one of the stable clusters, they could be perfectly (100 %) predicted from two acoustic features: the *zero-crossing rate*, reflecting the repetition rate of the signals, and the *standard deviation of the fundamental frequency* (see the lower panel of Figure 5). Accordingly, the imitations of liquids had a lower zero crossing rate, and a higher standard deviation of the fundamental frequency. This latter property is interesting, because it might be related to the short-term pitch glides characteristic of bubbles (Rocchesso & Fontana, 2003). Plotting the locations of the nine inconsistently clustered imitations drew a coherent picture: the imitations of liquids fell with the liquids, and the electrical appliances with the electrical appliances.

Division D/E: liquids and solids. Divisions D and E separate the three categories $\mathcal{L}_2$, $\mathcal{S}'$ and $\mathcal{S}''$. We first ran the decision tree on the nineteen imitations that were consistently clustered. In this case the imitations in the three clusters $\mathcal{L}_2$, $\mathcal{S}'$ and $\mathcal{S}''$ were perfectly discriminated (100 %) with two features (see the upper panel of Figure 6). The *standard deviation of the spectral centroid* distinguished $\mathcal{S}'$ from $\mathcal{L}_2$ and $\mathcal{S}''$. These features captured the variations of timbre across time occurring for the imitations in $\mathcal{L}_2$ and $\mathcal{S}''$ (whereas imitations in $\mathcal{S}'$ had a steadier timbre). Then the *zero-crossing rate*, reflecting the repetition rate of the signals, discriminated between the imitations with a low pitch in $\mathcal{L}_2$ and the sounds with a higher pitch in $\mathcal{S}$. It must therefore be noted that these three categories were discriminated on the basis of features that were more difficult to associate with how the event creating the sounds would have structured the sounds. Drawing the four imitations that were not consistently clustered showed that three of these sounds were close to the cluster $\mathcal{L}_2$.

Division F/G: gases G1, G2 and G3. The divisions F and G separate the imitations of the three gases $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$ + $\mathcal{X}$ (one cluster $\mathcal{X}$ is at this level associated with $\mathcal{G}_3$). Two features allowed for the perfect discrimination (100 %) of the 20 imitations consistently clustered in these categories (see the middle panel of Figure 6): the effective duration of the imitations, separating the shortest imitations in $\mathcal{G}_1$ from the other ones, and the standard deviation of the fundamental frequency, separating the imitations with a steady whistle in $\mathcal{G}_2$ from the other ones. One of the two sounds that were not well aggregated fell logically with the neighbor clusters $\mathcal{G}_2$ and $\mathcal{X}$.

Division H. The division H separates the cluster of imitations $\mathcal{G}_3$ from the composite cluster $\mathcal{X}$. The separation can be predicted from two features (see the lower panel of Figure 6). The standard deviation of the aperiodicity and the range of the fundamental frequency. One sound was misclassified over the nine considered here (88.9 %). However, it must be noted that the division is very low in the dendrogram, which indicates that, overall, these sounds were considered as similar.

Applying the imitation features to the original sounds

Because the imitations corresponding to a same referent sound were grouped together, and were characterized by few number of common features, it is now tempting to observe if these features could be used to classify the referent sounds. To conduct such a test, we computed the same acoustic features for the referent sounds, and observed the location of the sounds in the feature space. However, only the first level of the categorization could be tested (gases vs. solids and liquids vs. electrical appliances): at finer level, either the categories of imitations corresponded to a single referent sound (and thus, the categorization is trivial), or the categories of imitations mixed different referent sounds. The upper panel of Figure 7 represents the referent sounds in the modulation amplitude/minimum aperiodicity space.

Ten of the twelve sounds were correctly classified simply by the combination of two features the modulation amplitude and the average fundamental frequency (adding more features would have been erring on the side of overfitting). This figure presents a number of similarities with the upper panels of Figure 5, though the exact locations of the boundaries between the classes are different. Here, the electrical appliances were defined by their low fundamental frequency, and the absence of modulation. Gases were characterized by their high fundamental frequency, and the absence of modulation. Solids and liquids were characterized by their modulation. The sounds L_2 and L_3 were misclassified, but so were their imitations. There is therefore no reason to expect them to be correctly classified.

Because of the few number of sounds used here, these results should be interpreted cautiously. To assess the generality of the results, we also drew the position of the other sounds in the clusters where the referent sounds originated (Figure 1). Only four of these eleven sounds were correctly classified (see the lower panel of Figure 7).

Discussion

The binary decision tree analysis showed that the consistent clusters of imitations resulting from the experimental classification could be predicted from a few binary rules based on a few meaningful acoustic features. These clusters were therefore coherent both in terms of the category of the referent sound events, and in terms of their acoustic properties. The imitations that imitate the same category of sound events were acoustically similar. It is also interesting to note that most of the imitations that could not be clustered with other imitations could in fact be grouped together with a coherent cluster if one considers only the features of our analysis. The fact that the listeners did not group them together with other sounds suggest that they might also have some idiosyncratic properties that distinguish them from the other sounds.

In most cases[6], the similarity of the vocal production overlapped with the perception of the event causing the referent sound. But the acoustic analysis of some clusters (in particular the imitations of sounds of solids) also highlighted clusters that group together sounds sharing common acoustic properties, but mixing different events.

Finally, the success of the features found for the imitations to predict the categorization of the corresponding referent sounds indicates that studying the classification of the vocal imitations may have the potential to understand the features that characterize a

[6] And one might assume that these are the cases of efficient imitations.

given category of sound events. However, the lack of generalization to sounds other than the precise referent sounds also suggests that the transposition of the vocal features to any sound might be not straightforward.

General discussion

The study reported in this article was motivated by the following observation: a speaker very often makes use of vocal imitations to describe what he or she has in mind when he or she wants to communicate a sound to someone else. We then asked two questions: do the vocal imitations of sounds allow a listener to recover the imitated sound? If yes, could we study vocal imitations to assess what acoustic features are necessary for sound event identification?

We have reported an experiment examining these questions. We studied the meaning conveyed by a set of non-conventional vocal imitations of everyday sounds, specifically created by a number of imitators and selected for the purpose of this study. The imitators vocalized a set of sounds identifiable at two levels of specificity: the specific source, or the type of sound production. Another set of subjects (listeners) were required to sort these vocal imitations on the basis of what they thought was imitated. The results showed that the categories of imitations created by the listeners in general corresponded to the referent sound. Some imitations of the same specific source were clustered together. Many others fell into clusters corresponding to the type of sound production. In fact, the referent sounds had been chosen precisely because they belonged to contrasting categories of different sounding events. Therefore, these results suggest that the vocal imitations conveyed enough information for the listener to recover at least the type of sound production to which the referent sound belonged; and even in some cases the exact specific source. However, not all the sound events were correctly recovered: in particular, the imitations of liquids were incorrectly classified. Instead, they were either clustered with other irrelevant imitations, or not consistently clustered. The analysis of these imitations suggested that the imitators were unable to successfully render the "liquid character" of these sounds with their voice (in particular the short chirps that are suspected to be characteristic of liquids). The first possible explanation might be that the imitators did not recognize the referent sound events, and therefore tried to vocally convey the wrong type of information. This seems unlikely, because these referent sound events were precisely selected on the basis of their easy identifiability. Another explanation might be that the acoustic features characteristic of liquids were difficult to render with the human vocal apparatus. Listening to the imitations of liquid sounds suggests that the imitators had instead emphasized some coarser features (in particular the temporal structure) that made the listeners confound them with other sounds presenting similar patterns. This failure of communicating the liquid identity therefore appears to have resulted at least in part from a physiological constraint: the impossibility to render at the same time a turbulent noisy component and tonal glides.

The most striking result of this study was that the resulting clusters of imitations appeared to be characterized by a few simple acoustic features. This intuition was confirmed by submitting the data to a very simple type of machine learning technique: a binary decision tree analysis. This analysis showed that the clustering of the imitations could be predicted from binary decisions based on a few of acoustic features. So it appears that the listeners have only used a limited number of simple acoustic features to cluster the

imitations. These features did not imply any complex characteristic but, apparent simple characteristics: continuous vs. rhythmic sounds, tonal vs. noisy, short vs. long, etc. These coarse features were sufficient for the listeners to recover the types of sound production. That these clusters corresponded to the categories of referent sound events implies that the imitators have probably chosen to vocally emphasize a few characteristics of the sounds that they believed would be sufficient for recognition of these categories, within the limits of what is vocally possible.

The question is then to know whether the imitators chose to emphasize these features because they were characteristic of the different categories of sound events *in general*, or because they were the most distinctive features that distinguished this *particular* set of sounds. The former explanation would imply that different psychologically relevant categories of sound events can be defined, in general, by a few invariant acoustic features. In fact, when we tried to use the features responsible for the clustering of imitations to predict the categorization of the referent sounds, we were fairly successful if we limited ourselves to the precise sounds that had been imitated. However, we failed to successfully predict the categorization of other sounds (sounds that had not been imitated) from the same categories. This suggests that, even if the vocal imitations successfully conveyed the features that enabled the listeners to successfully distinguish the imitations of the different sound events, these features were probably not sufficient to characterize these categories in general.

On the basis of these observations, we can now propose an interpretation of the imitators' strategy. Required to vocally imitate a set of sounds that clearly fell into four categories of sound events, they picked up a few of features: those that they could easily render vocally, and that they thought could maximally distinguish these categories. But these features were only distinctive of this particular set of sounds: the imitators selected information on the basis of the task and the set of sounds. For instance, they used the duration of the sounds to distinguish some of the categories. If this was a clever choice for our particular set of sounds, the duration of the sounds cannot be thought as characteristic of certain categories of sound events in general. Such an "opportunistic" behavior can also be observed in free categorization experiments: presented with a number of sounds, listeners tend to pick up the acoustic information that would allow them to form not too many, nor too few categories (an effect that is not unlike what Parducci & Wedell, 1986 have reported for category scales. Notice that McAdams et al. (2010) also reported that listeners could pick up different kind of information from a same set of sounds, depending on the task). This also suggests that listeners do not have definitively fixed low-dimensional representations of sounds: they can use different aspects of the sounds to adapt their behavior to the task.

The main purpose of this report was to explore whether studying vocal imitations of sounds could help us understand the auditory features necessary for the identification of different sound events. The answer is mixed. In the simple case that we have reported here, studying the vocal imitations showed the features of the sounds that the imitators chose to convey to the listeners. But because these features were probably useful only locally, for our particular set of sounds, we did not really learn anything about what characterizes these sound events in general. But these results nevertheless show that the process of vocally imitating sounds consists of picking up certain characteristics of the sounds, and vocally conveying these features to the listeners, probably in an emphasized fashion. As such,

studying vocal imitations would help us understand what imitators think is important for recognition, within the limits of what is vocally possible. More sophisticated designs are now necessary to investigate the questions more thoroughly.

A potential limitation of this work is that we have applied the features found for the vocal imitations to the referent sounds without any adaptation. We did not take into account the important differences between vocal and non-vocal sounds. It is in fact very likely that imitators did not try to exactly mimic the characteristics of the sounds, mainly because not every acoustic feature is reproducible by the human vocal apparatus. Some kind of transcription probably occurred: for instance, imitators did not try to exactly reproduce the sounds of the motors of the kitchen appliances: rather they used voiced and unvoiced vocal sounds to signify the presence or absence of a motor. Further work is now needed to understand how imitators use different places of articulation to render different kinds of mechanical sound events.

Another interesting perspective will be to study vocal imitations during conversations. For instance, future studies could ask: In which cases do speakers use vocal imitations and onomatopoeias? How is the use of imitations related to the identifiability of the sound source? To the availability of a relevant vocabulary? To the physiological constraints of the human voice? Another relevant question is that of the effectiveness of vocal imitations to communicate a sound in the more ecological context of a conversation: for sounds that are difficult to describe with words, how effective can speakers be, when they are allowed or forbidden to use vocal imitations?

Besides fundamental questions, a better understanding of these processes will potentially lead to many practical applications in audio content analysis or in sound synthesis: search-by-similarity, query-by-example, automatic classification, etc. More specifically, studying sound event identification and vocal imitations is expected to inform the development of *cartoonification* (Rocchesso et al., 2003). The advantages of such a technique are that it renders the information clearer, and more effective, while reducing the computational costs. Using vocal imitation to control of sound synthesis is another promising approach (Ekman & Rinott, 2010). The development of all these applications will, particularly, require us to precisely understand how speakers use different vocal sounds and manners of articulation to communicate specific sound events (the production of vocal imitations), and how listeners "decode" the vocal productions to recover the referent sound events and sources (the perception and cognition of these imitations).

References

Abdi, H., Valentine, D., Chollet, S., & Chrea, C. (2007). Analyzing assessors and products in sorting tasks: Distatis, theory and applications. *Food quality and preference*, *18*(4), 627-640.

Aldrich, K. M., Hellier, E. J., & Edworthy, J. (2009). What determines auditory similarity? the effect of stimulus group and methodology. *Quarterly Journal of Experimental Psychology*, *62*(1), 63-83.

Ballas, J. A. (1993). Common factors in the identification of an assortment of brief everyday sounds. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(2), 250-267.

Ballas, J. A., & Howard, J. H. (1987). Interpreting the language of environmental sounds. *Environment and Behavior*, *19*(1), 91-114.

Ballas, J. A., & Mullins, T. (1991). Effect of context on the identification of everyday sounds. *Human Performance*, *4*(3), 199-219.

Boersma, P., & Weenink, D. (2009). *Praat: doing phonetics by computer (version 5.1.05).* (Computer program. Retrieved May 1, 2009, from http://www.praat.org/)

Cabe, P. A., & Pittenger, J. B. (2000). Human sensitivity to acoustic information from vessel filling. *Journal of experimental psychology: human perception and performance*, *26*(1), 313-324.

Carello, C., Anderson, K. L., & Kunkler-Peck, A. J. (1998, May). Perception of object length by sound. *Psychological science*, *9*(3), 211-214.

Carello, C., Wagman, J. B., & Turvey, M. T. (2005). Acoustic specification of object property. In J. D. Anderson & B. Fisher Anderson (Eds.), *Moving image theory: ecological considerations* (p. 79-104). Carbondale, IL: Southern Illinois University Press.

Cummings, A., Čeponienė, R., Katoma, A., n, A. P. S., Townsend, J., & Dick, F. (2006). Auditory semantic networks for words and natural sounds. *Brain research*, *115*, 92-107.

de Cheveigné, A., & Kawahara, H. (2001). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, *111*(4), 1917-1930.

Ekman, I., & Rinott, M. (2010). Using vocal sketching for designing sonic interactions. In *DIS '10: Proceedings of the 8th ACM conference on designing interactive systems* (pp. 123–131). New York, NY, USA: ACM.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometris*, *29*(4), 751-760.

Gaver, W. W. (1993a). How do we hear in the world? Explorations in ecological acoustics. *Ecological Psychology*, *5*(4), 285-313.

Gaver, W. W. (1993b). What do we hear in the world? An ecological approach to auditory event perception. *Ecological Psychology*, *5*(1), 1-29.

Gillet, O., & Richard, G. (2005). Drum loops retrieval from spoken queries. *Journal of Intelligent Information Systems*, *24*(2/3), 160-177.

Giordano, B. L., & McAdams, S. (2006, February). Material identification of real impact sounds: effect of size variation in steel, glass, wood and plexiglass plates. *Journal of the Acoustical Society of America*, *119*(2), 1171-1881.

Giordano, B. L., McAdams, S., & Rocchesso, D. (2010). Integration of acoustical information in the perception of impacted sound sources: the role of information accuracy and exploitability. *Journal of experimental psychology: human perception and performance*, *36*(2), 462-476.

Giordano, B. L., McDonnell, J., & McAdams, S. (2010). Hearing living symbols and nonliving icons: category specificities in the cognitive processing of environmental sounds. *Brain and cognition*, *73*, 7-19.

Grassi, M. (2005). Do we hear size or sound? Balls dropped on plates. *Perception and Psychophysics*, *67*(2), 274-284.

Handel, S. (1989). Listening: an introduction to the perception of auditory events. In (p. 219-274). Cambridge, MA: The MIT Press.

Handel, S. (1995). Timbre perception and auditory object identification. In B. C. J. Moore

(Ed.), *Hearing. Handbook of perception and cognition* (Second ed., p. 425-461). Academic Press.

Hashimoto, T., Usui, N., Taira, M., Nose, I., Haji, T., & Kojima, S. (2006). The neural mechanism associated with the processing of onomatopoeic sounds. *Neuroimage*, *31*, 1762-1170.

Heller, L. M., & Wolf, L. (2002). When sound effects are better than the real thing. In *Proceedings of the 143rd ASA meeting.* Pittsburgh, PA.

Houben, M. M. J., Kohlrausch, A., & Hermes, D. J. (2004). Perception of the size and speed of rolling balls by sound. *Speech communication*, *43*, 331-345.

Houix, O., Lemaitre, G., Misdariis, N., Susini, P., & Urdapilleta, I. (2011). A lexical analysis of environmental sound categories. *Manuscript submitted for publication*.

Howard, J. H., & Ballas, J. A. (1980). Syntactic and semantic factors in the classification of nonspeech transient patterns. *Perception and Psychophysics*, *28*(5), 431-439.

Ishihara, K., Nakatani, T., Ogata, T., & Okuno, H. G. (2004). Automatic sound-imitation word recognition from environmental sounds focusing on ambiguity problem in determining phonemes. In C. Zhang, H. W. Guesgen, & W.-K. Yeap (Eds.), *Pricai* (Vol. 3157, p. 909-918). Springer.

Ishihara, K., Tsubota, Y., & Okuno, H. G. (2003). Automatic transcription of environmental sounds into sound-imitation words based on japanese syllable structure. In *Proceedings of eurospeech 2003* (p. 3185-3188). Geneva, Swirtzerland.

Iwasaki, N., Vinson, D. P., & Vigliocco, G. (2007). What do English speakers know about *gera-gera* and *yota-yota*? A cross-linguistic investigation of mimetic words for laughing and walking. *Japanese-language education around the globe*, *17*, 53-78.

Klatzky, R. L., Pai, D. K., & Krotkov, E. P. (2000). Perception of material from contact sounds. *Presence*, *9*(4), 399-410.

Kunkler-Peck, A. J., & Turvey, M. T. (2000). Hearing shape. *Journal of Experimental psychology: human perception and performance*, *26*(1), 279-294.

Lakatos, S., McAdams, S., & Caussé, R. (1997). The representation of auditory source characteristics: simple geometric forms. *Perception & psychophysics*, *59*(8), 1180-1190.

Lass, N. J., Eastham, S. K., Parrish, W. C., Sherbick, K. A., & Ralph, D. M. (1982). Listener's identification of environnmental sounds. *Perceptual and Motor Skills*, *55*, 75-78.

Lass, N. J., Eastham, S. K., Wright, T. L., Hinzman, A. H., Mills, K. J., & Hefferin, A. L. (1983). Listener's identification of human-imitated sounds. *Perceptual and Motor Skills*, *57*, 995-998.

Lass, N. J., Hinzman, A. H., Eastham, S. K., Wright, T. L., Mills, K. J., Bartlett, B. S., et al. (1984). Listener's discrimination of real and human-imitated sounds. *Perceptual and Motor Skills*, *58*, 453-454.

Lemaitre, G., Dessein, A., Aura, K., & Susini, P. (2009). Do vocal imitations enable the identification of the imitated sounds? In *Proceedings of the 8th annual Auditory Perception, Cognition and Action Meeting (APCAM 2009).* Boston, MA.

Lemaitre, G., & Heller, L. M. (2010). Action verbs are the most accessible level of sound event description. In *Proceedings of the 9th annual Auditory Perception, Cognition and Action Meeting (APCAM 2010).* St. Louis, MO.

Lemaitre, G., & Heller, L. M. (2011). Auditory perception of material is fragile, while action is strikingly robust. *Manuscript submitted for publication*.

Lemaitre, G., Houix, O., Misdariis, N., & Susini, P. (2010). Listener expertise and sound identification influence the categorization of environmental sounds. *Journal of Experimental Psychology: applied*, *16*(1), 16-32.

Lemaitre, G., Susini, P., Winsberg, S., Letinturier, B., & McAdams, S. (2007). The sound quality of car horns: a psychoacoustical study of timbre. *Acta Acustica united with Acustica*, *93*(3), 457-468.

Lemaitre, G., Susini, P., Winsberg, S., Letinturier, B., & McAdams, S. (2009). The sound quality of car horns: Designing new representative sounds. *Acta Acustica united with Acustica*, *95*(2), 356-372.

Lutfi, R. A., & Oh, E. L. (1997, December). Auditory discrimination of material changes in a struck-clamped bar. *Journal of the Acoustical Society of America*, *102*(6), 3647-3656.

Marozeau, J., Cheveigné, A. de, McAdams, S., & Winsberg, S. (2003). The dependency of timbre on fundamental frequency. *Journal of the Acoustical Society of America*, *114*(5), 2946-2957.

McAdams, S., Chaigne, A., & Roussarie, V. (2004, March). The psychomechanics of simulated sound sources: material properties of impacted bars. *Journal of the Acoustical Society of America*, *115*(3), 1306-1320.

McAdams, S., Roussarie, V., Chaigne, A., & Giordano, B. L. (2010). The psychomechanics of simulated sound sources: material properties of impacted thin plates. *Journal of the Acoustical Society of America*, *128*, 1401-1413.

Nakano, T., & Goto, M. (2009). Vocalistener: a singing-to-singing synthesis system based on iterative parameter estimation. In *Proceedings of the Sound and Music Computing (SMC) conference 2009.* Porto, Portugal.

Nakano, T., Ogata, J., Goto, M., & Hiraga, Y. (2004). A drum pattern retrieval method by voice percussion. In *Proceedings of the* 5^{th} *International Conference on Music Information Retrieval (ISMIR 2004)* (p. 550-553). Barcelona, Spain.

Newman, F. (2004). *Mouthsounds: How to whistle, pop, boing and honk for all occasions... and then some*. Workman Publishing Company.

Oswalt, R. L. (1994). Inanimate imitatives. In L. Hinton, J. Nichols, & J. Ohala (Eds.), *Sound symbolism.* Cambridge University Press.

Parducci, A., & Wedell, D. H. (1986). The category effect with rating scales: number of categories, number of stimuli and method of presentation. *Journal of Experimental Psychology: human perception and performance*, *12*(4), 496-516.

Patel, A., & Iversen, J. (2003). Acoustical and perceptual comparison of speech and drum sounds in the North India tabla tradition: an empirical study of sound symbolism. In *Proceedings of the* 15^{th} *international congress of phonetic sciences.* Barcelona, Spain.

Peeters, G. (2004). *A large set of audio features for sound description (similarity and classification) in the CUIDADO project* (Cuidado Projet report). Paris, France: Institut de Recherche et de Coordination Acoustique Musique (IRCAM).

Pharies, D. A. (1979). *Sound symbolism in the Romance languages*. Unpublished doctoral dissertation, University of Columbia, Berkeley.

Proctor, M., Nayak, K., & Narayanan, S. (2010). Para-linguistic mechanisms of production in human 'beatboxing': a real-time mri study. In *Proceedings of InterSinging 2010.*

Tokyo, Japan.

Rhodes, R. (1994). Aural images. In L. Hinton, J. Nichols, & J. Ohala (Eds.), *Sound symbolism.* Cambridge University Press.

Rocchesso, D., Bresin, R., & Fernström, M. (2003, April-june). Sounding objects. *IEEE Multimedia*, *10*(2), 42-52.

Rocchesso, D., & Fontana, F. (Eds.). (2003). *The sounding object.* Florence, Italy: Ediziono di Mondo Estremo.

Sobkowiak, W. (1990). On the phonostatistics of English onomatopoeia. *Studia Anglica Posnaniensia*, *23*, 15-30.

Strange, W., & Shafer, V. (2008). Speech perception in second language learners: the re-education of selective perception. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (p. 153-192). Philapelphia, PA: John Benjamin Publishing Company.

Sundaram, S., & Narayanan, S. (2006). Vector-based representation and clustering of audio using onomatopoeia words. In *Proceedings of the American Association for Artificial Intelligence (AAAI) symposium series.* Arlington, VA.

Sundaram, S., & Narayanan, S. (2008). Classification of sound clips by two schemes: using onomatopeia and semantic labels. In *Proceedings of the IEEE conference on multimedia and expo (ICME)* (p. 1341-1344). Hanover, Germany: IEEE.

Susini, P., McAdams, S., & Winsberg, S. (1999). A multidimensional technique for sound quality assessment. *Acustica united with Acta Acustica*, *85*, 650-656.

Takada, M., Fujisawa, N., Obata, F., & Iwamiya, S. (2010). Comparisons of auditory impressions and auditory imagery associated with onomatopoeic representations for environmental sounds. *EURASIPJournal on Audio, Speech, and Music Processing*. (Article ID 674248)

Takada, M., Tanaka, K., & Iwamiya, S. (2006). Relationships between auditory impressions and onomatopoeic features for environmental sounds. *Acoustic Science and Technology*, *27*(2), 67-79.

Takada, M., Tanaka, K., Iwamiya, S., Kawahara, K., Takanashi, A., & Mori, A. (2001). Onomatopeic features of sounds emitted from laser printers and copy machines and their contributions to product image. In *Proceedings of the international conference on acoustics ICA 2001.* Rome, Italy.

Troubetzkoy, N. S. (1949). *Principe de phonologie.* Paris, France: Librairie Klincksieck.

Vanderveer, N. J. (1979). *Ecological acoustics: human perception of environmental sounds.* Unpublished doctoral dissertation, Cornell University.

Warren, W. H., & Verbrugge, R. R. (1984). Auditory perception of breaking and bouncing events: a case study in ecological acoustics. *Journal of Experimental Psychology: Human perception and performance*, *10*(5), 704-712.

Wildes, R. P., & Richards, W. A. (1988). Recovering material properties from sound. In W. A. Richards (Ed.), *Natural computation* (p. 356-363). Cambridge, MA. London, England: A Bradford book. The MIT press.

Wright, P. (1971). Linguistic description of auditory signals. *Journal of applied psychology*, *55*(3), 244-250.

Żuchowski, R. (1998). Stops and other sound-symbolic devices expressing the relative length of referent sounds in onomatopoeia. *Studia Anglica Posnaniensia*, *33*, 475-485.

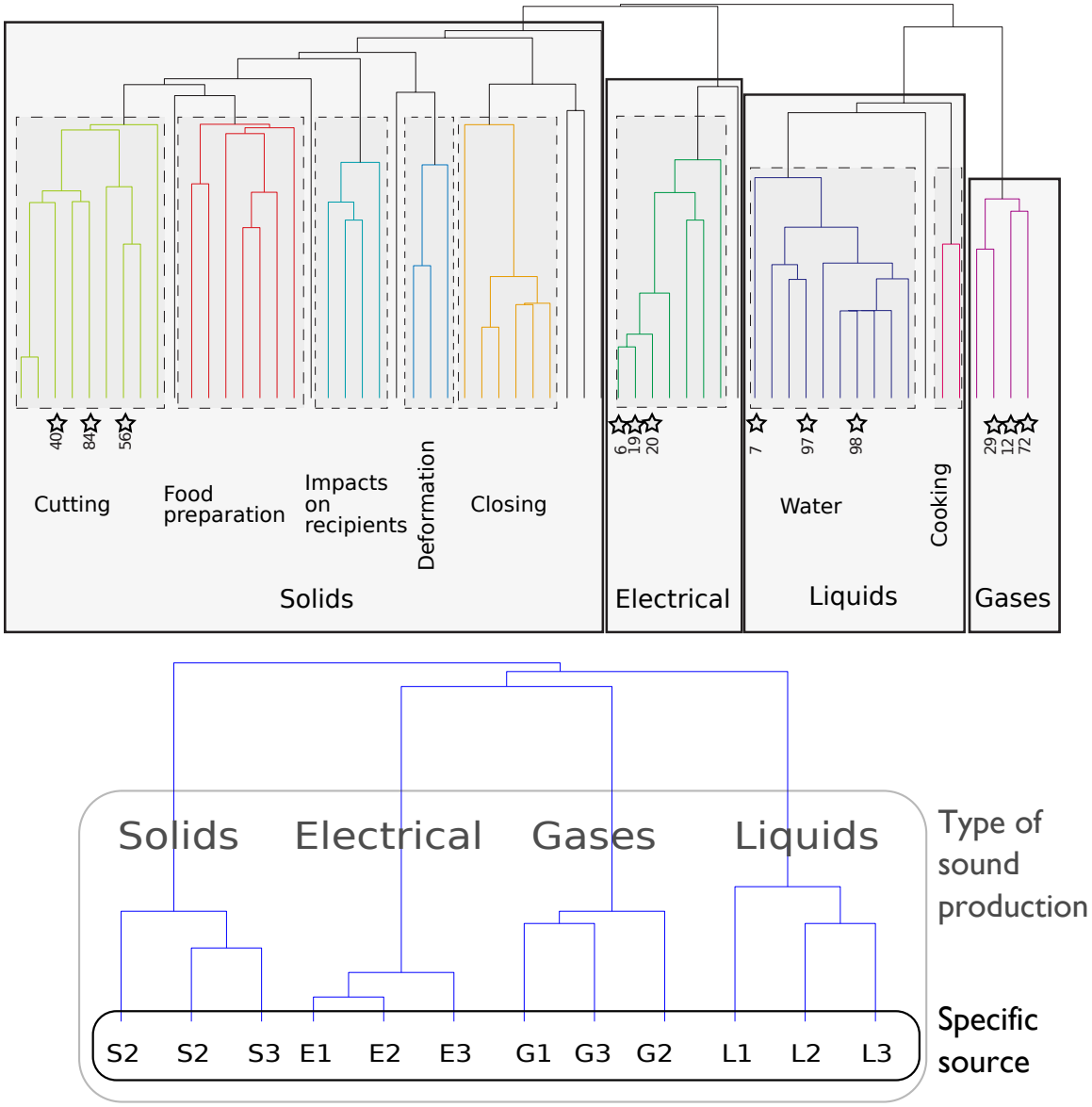

Figure 1 Hierarchical tree representation of the categorization of the initial kitchen sounds. The stars correspond to sounds, the imitations of which were finally selected for the subsequent experiments. The upper panel is adapted from Houix (2010). The lower panel corresponds to the same cluster analysis as used with the imitations (see the text for details), applied only to the 12 referent sounds. This panel shows the two levels of specificity for the description of each sound: the type of sound production, and the specific source.

Table 1: List of the referent sounds to be imitated. The confidence values were measured in Lemaitre et al. (2010). The confidence scale varied from 0 ("I do not know at all what has caused the sound") to 8 ("I perfectly identify the cause of the sound"). The index in parentheses makes reference to this latter article.

#	Description	Max. level (dB)	Duration (s)	Confidence
	Solid objects (cutting)			
S1 (40)	Seven repetitions of a dented knife cutting through the crust of a loaf of bread.	61	1.2	6.68
S2 (56)	Eight twisting sounds of a lid being screwed back on a container.	65	2.9	5.31
S3 (84)	Eight impacts of a knife cutting through a vegetable to a cutting board placed underneath.	67	6.2	7.05
	Electrical sounds (machines)			
E1 (6)	Dishwasher on. Low steady hum.	66	3.0	4.84
E2 (19)	Food processor. Low steady hum.	72	3.5	3.89
E3 (20)	Mixer on. Low steady hum.	72	2.8	6.21
	Liquids (water)			
L1 (7)	Coffee maker with filter on. Two gusts of coffee passing through the filter creating two bursts of noise.	62	6.9	6.63
L2 (97)	Three drips falling in a container. The sound consists of three rapid frequency sweeps.	64	3.6	7.42
L3 (98)	Water running in a sink. Long steady broadband noise	67	5.0	7.00
	Gases			
G1 (12)	Striking and igniting a match. The sound begins with the scratch of the match, followed by the match ignition.	62	1.1	6.21
G2 (29)	Gas open on a stove. Constant broadband noise with a strong resonance if the lower frequencies.	42	3.6	6.26
G3 (72)	Spray from an aerosol. Broadband noise with a modulation similar to a time-varying comb filter.	68	3.7	5.31

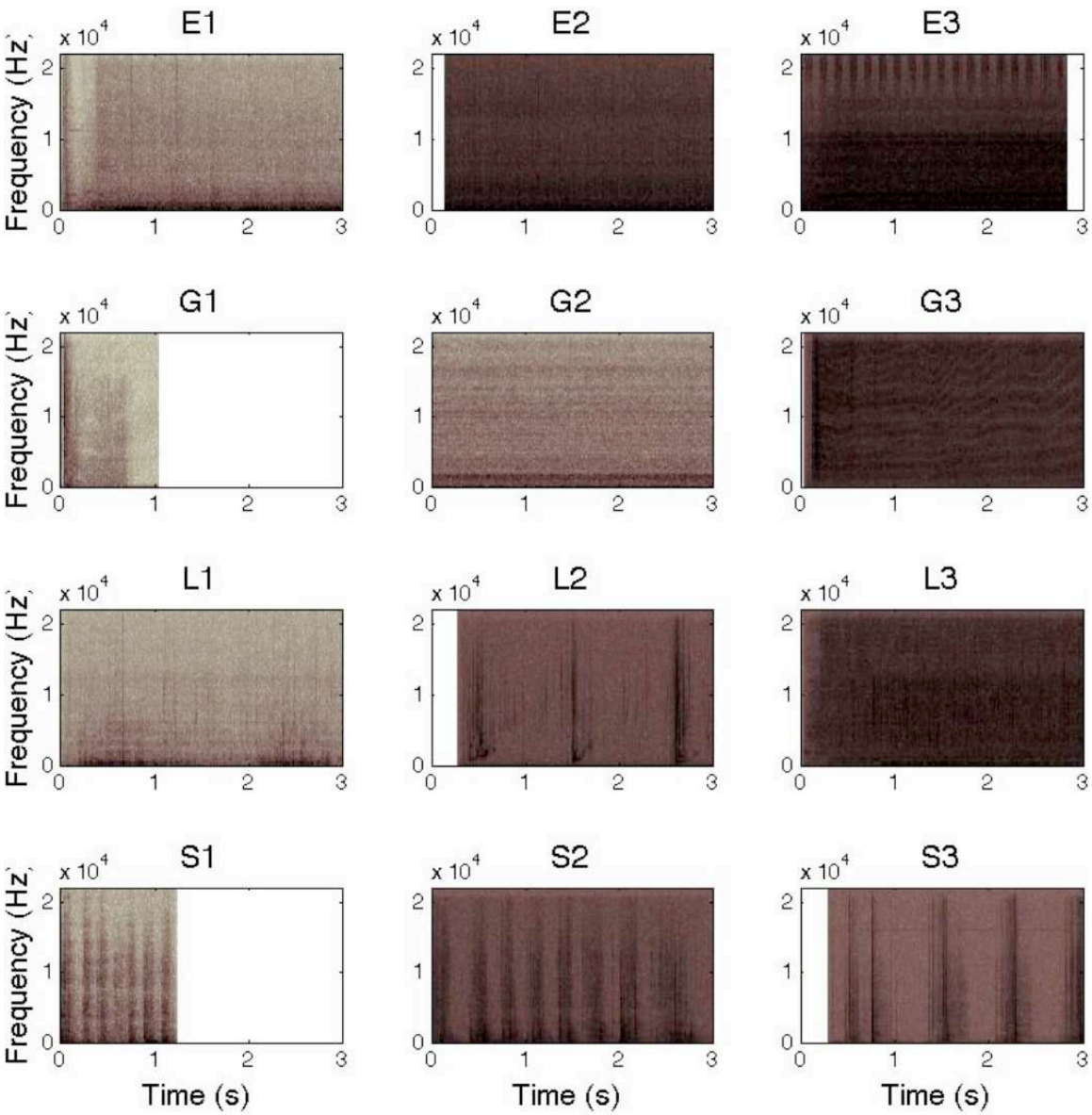

Figure 2 Spectrograms of the 12 referent sounds.

Table 2: Phonetic transcription of the 72 imitations of the 12 referent sounds vocalized by 6 imitators (three men - lower panel - and three women - upper panel). The phonetic symbols refer to French phonemes (International Phonetic Alphabet); no distinction is here made between /ʀ/ and /ʁ/. These transcription are approximative (and sometimes impossible), because in many cases the imitators used sounds that were outside of the range of sounds used in French.

Referent sound	W1	W2	W3
E1	[brəː]	[kvu]	[tutuduˈtuduː]
E2	[tʀuː]	[pʀːʀœ̃ː]	[pʀwərwərwərwərwərwərwərwə]
E3	[dʒəː]	[aʀœ̃ː]	[tərœːməː]
G1	[tuːf]	[pfyːfə]	[truˈtu]
G2	[fəː]	[ʀə]	[fuː]
G3	[piəwisə]	[pfəː]	[tədu-ʃyʃyʃyʃyʃy]
L1	[trərərufʀufʀuf]	[fʀurərərəː]	[pʀuffufufufəːpʀuf]
L2	[pələwˈ-pləwˈ-pləw]	[ptui-tui-tui]	[plwik-pluk-pluk]
L3	[pwu-pʒu-ʒiː]	[ʃiːuːfuː]	[tʀərərə-rərərə-fuː]
S1	[frətətʀətʀətʀətʀə]	[fʀkʀəkʀəkʀəkʀəkʀə]	[fʀut-fʀutfʀut-fʀutfʀut]
S2	[tum-tum-tum-tum]	[ʀik-kʀəkʀəkʀəkʀəkʀəkʀə]	[ptʃykˈtudu-kʃyktʃyktʃyktʃyk]
S2	[ntɔk-ntɔk-ntuk-ntuk]	[kɥət-kɥət-kɥət-kɥət]	[futəˈtʀutʀutʀutʀutʀut]

Referent sound	M1	M2	M3
E1	[wəːwəː]	[dvuːvuːvuːvuːvuːv]	[vuː]
E2	[ʀəwu-ʀəwu-ʀəwu-ʀəwu]	[brəː]	[trrrrrrrəː]
E3	[ʀuː]	[vəː]	[ʀɛ̃ː]
G1	[tʃuːu]	[pfʃiː]	[tutˈfuː]
G2	[sː]	[ʃəː]	[fyː]
G3	[tʃufəːit]	[tʃiː]	[psijyː]
L1	not transcriptable	[kʀʀuːʀuːʀurəː]	[kʀɔːkˈkʀɔːkkʀɔːk]
L2	[pʀuk-pluk-plup]	[ʃviː-puk-puk]	[klwik-pa-pa]
L3	[ʀəʃʃyːʃyʃyː]	[ʃyː]	[tʃitə-ʃwauaːuːiː]
S1	[ʀərərərərərərə]	[fʀuʀu-ʀu-ʀu]	[tœˈtœˈtœ-tœˈtœˈtœˈtœ]
S2	[ʀərərərərərə]	[fʀəfʀəfəfʀəfʀə]	[tr-tr-tr-tr]
S3	[kwupkwupwupwup]	[pu-pu-pu-pu-pu-pu]	[fut-fut-fut-fut-fut]

Table 3: Some examples of descriptions of imitations provided by the participants, sorted into different kinds of similarity. Most of the descriptions fell into the categories "causal" and "semantic".

Type of similarity	Referent sound	Examples of description
Causal / Semantic	S1 (cutting bread)	"Cutting food"
	L2 (drips in a container)	"Water dripping"
	E3 (mixer)	"All kinds of drilling machines, food processors"
Acoustic	L2 (drips in a container)	"Loud and rhythmic sounds"
	S2 (cutting vegetables)	"Short and repeated sounds"
	E1 (dishwasher)	"Loud and continuous sound"
Hedonic	S1 (cutting bread)	"Very aggressive, catches attention"
	G1 (striking a match)	"Suffering"
	E3 (mixer)	"Mentions the comfort"
Vocal production	L3 (water running in a sink)	"Throat noises"
	S3 (cutting food)	"Expiration with a whistle on the tongue"
	L2 (drips in a container)	"With the lips"

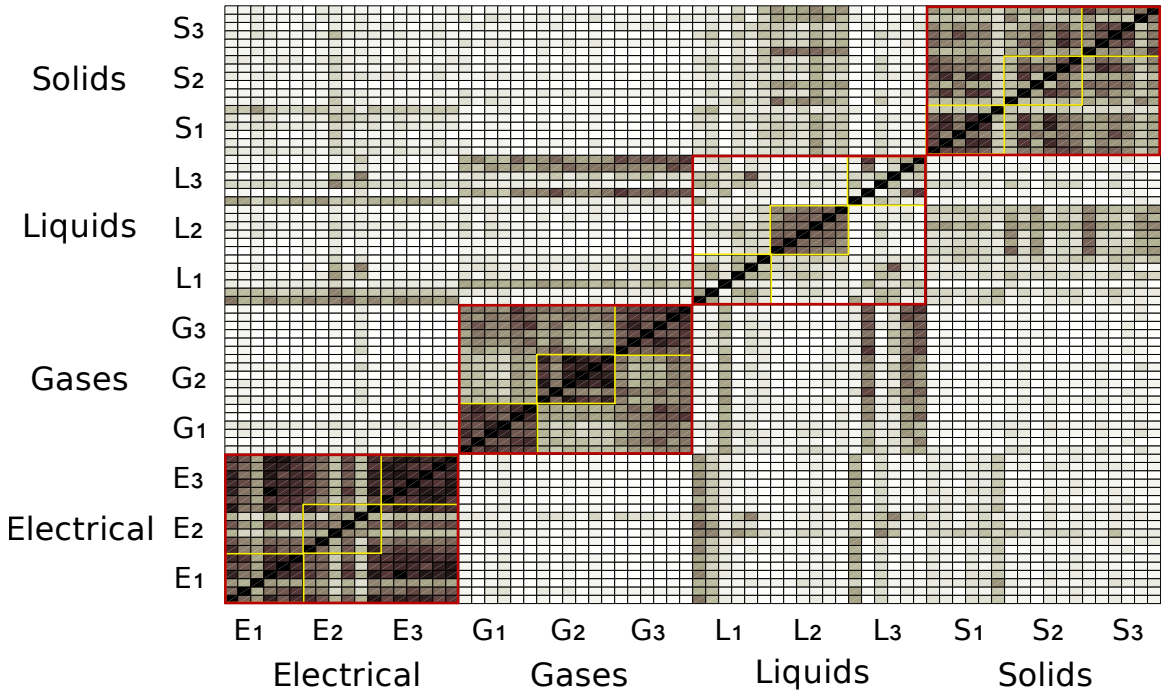

Figure 3 Confusion matrix: the color of each cell represents the number of participants who placed the two sounds in the same group.

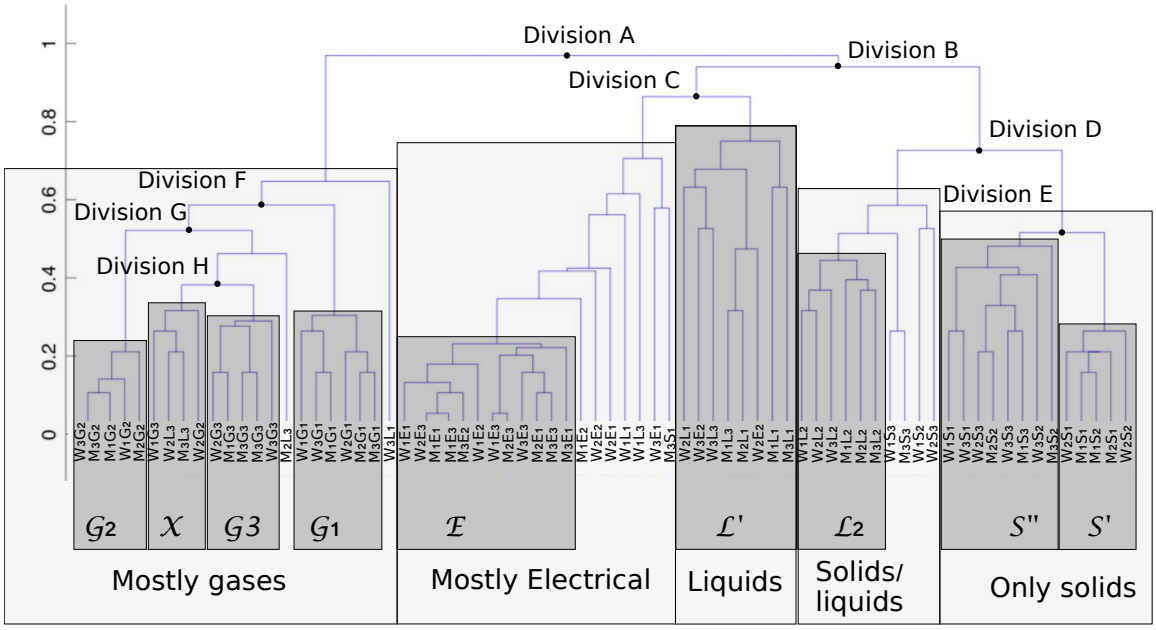

Figure 4 Dendrogram of the vocal imitations. The first letter of the index of each imitations describes the imitator (M stands for man, and W for woman), whereas the second letter describes the category of imitated sound (L is for liquid, S for solid, M for machine and G for gases), see Table 1.

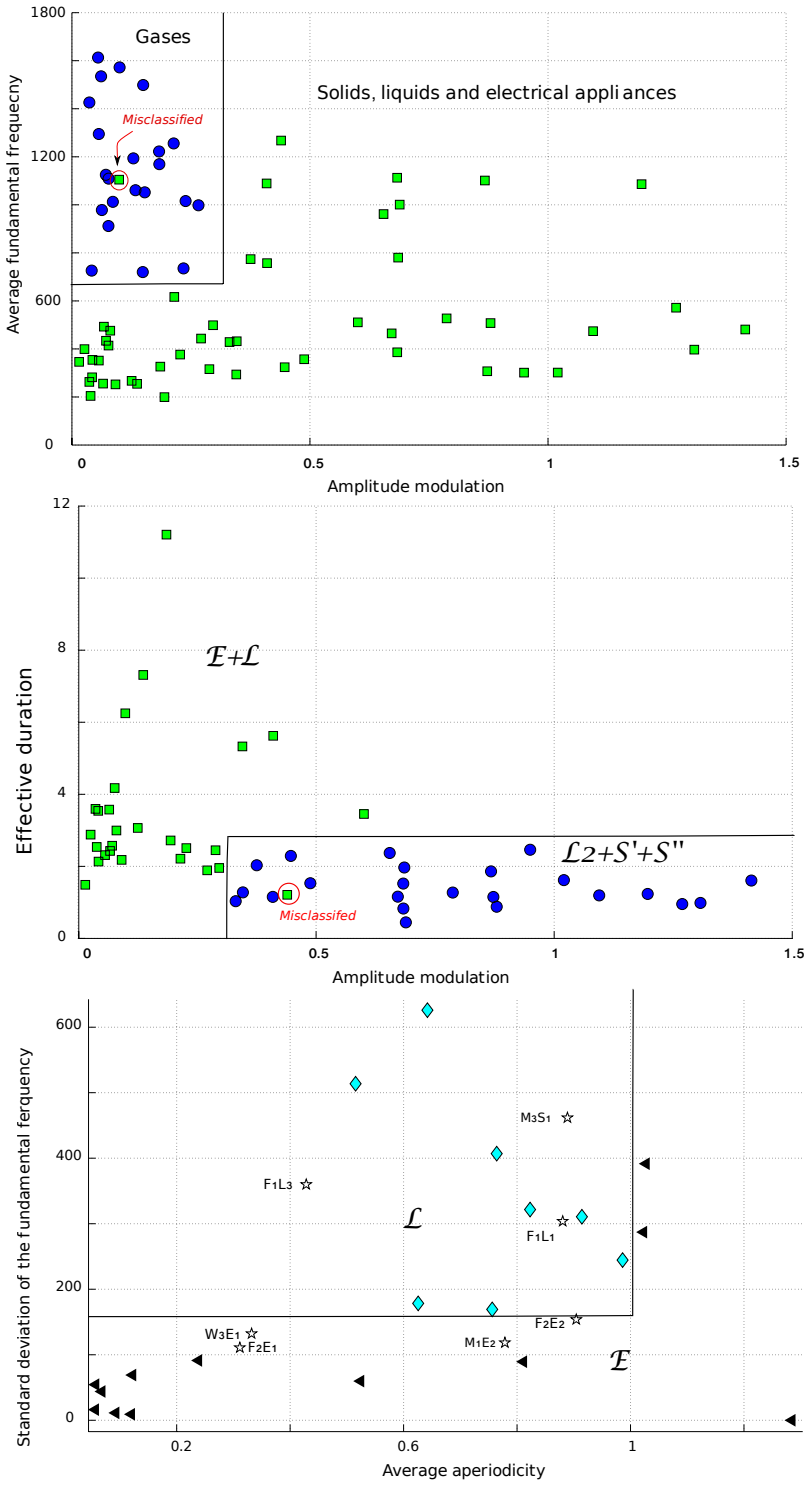

Figure 5 Clustering of the vocal imitations by the binary decision tree analysis. The upper panel corresponds to Division A in the dendrogram of imitations, separating the imitations of gases from all the other imitations. The middle panel represents Division B, and the lowest panel Division C. In these three panels, the filled symbols correspond to imitations that were consistently clustered, the empty symbols to imitations that could not be grouped together with any stable cluster.

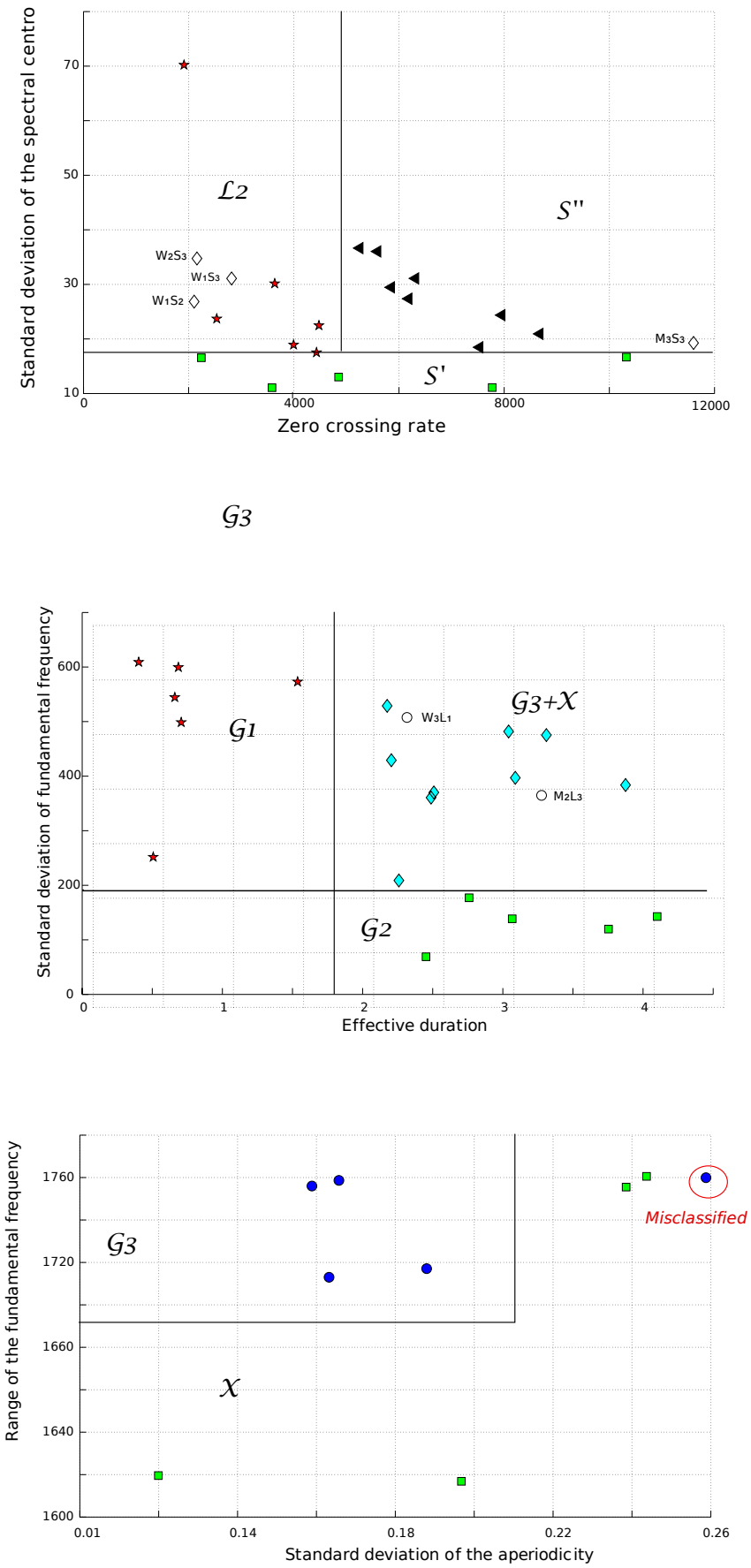

Figure 6 Clustering of the vocal imitations by the binary decision tree analysis. The upper panel corresponds to Divisions D and E, the middle panel represents Divisions F and G, and the lowest

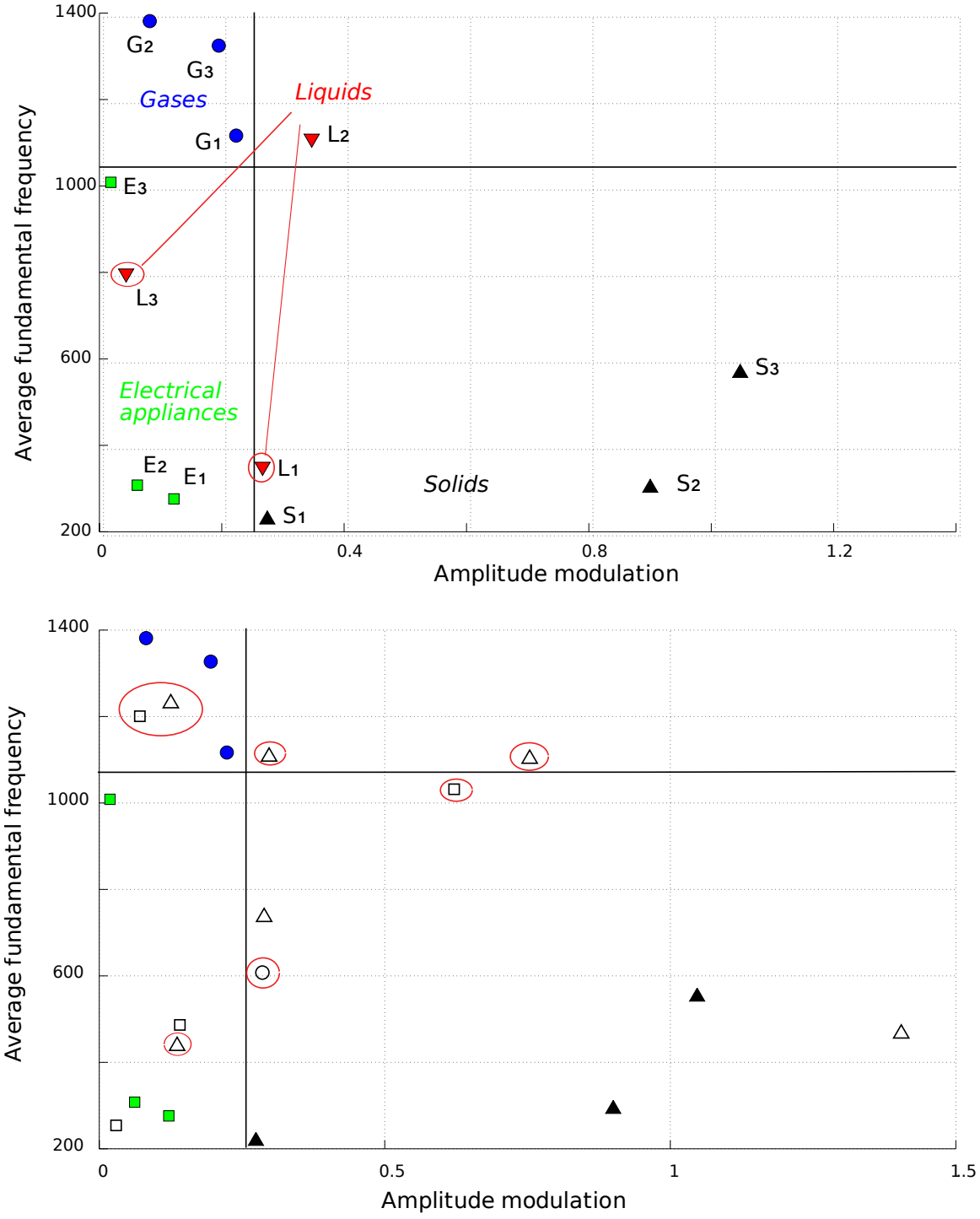

Figure 7 Upper panel: clustering of the referent sounds based on the features found for the vocal imitations. Lower panel: the sounds in the same categories of the original sounds have been added (empty symbols).

Appendix A
A. Identifying potential outliers in the classification experiment

To highlight potential differences between the participants' answers, we used a method inspired by Abdi, Valentine, Chollet, and Chrea (2007). It consists of computing a measure of pairwise similarity between the individual classifications, and adding some randomly generated artificial classifications in order to detect potential outliers. For each participant p, the results of the classification were encoded in a $n \times n$ matrix $\mathbf{D}_p$, called a *distance matrix*, such that:

$$d_{ij} = \begin{cases} 0 & \text{if sounds } i \text{ and } j \text{ were grouped together;} \\ 1 & \text{other.} \end{cases} \tag{1}$$

A measure of pairwise similarity The R_V coefficient (Escoufier, 1973) is a measure of similarity between two symmetric matrices $\mathbf{X}$ and $\mathbf{Y}$, and can be used as a measure of pairwise similarity between individual classifications:

$$R_V(\mathbf{X}, \mathbf{Y}) = \frac{\operatorname{trace}(\mathbf{X}\mathbf{Y}^T)}{\sqrt{\operatorname{trace}(\mathbf{X}\mathbf{X}^T)\operatorname{trace}(\mathbf{Y}\mathbf{Y}^T)}} \tag{2}$$

Following Abdi et al. (2007), the R_V coefficient is computed between the individual normalized (with respect to the spectral radius) *cross-product matrices*. The cross-product matrix $\tilde{\mathbf{S}}_p$ for participant p is given by:

$$\tilde{\mathbf{S}}_p = -\frac{1}{2}\mathbf{C}\mathbf{D}_p\mathbf{C}^T \tag{3}$$

where $\mathbf{D}_p$ is the distance matrix of participant p. The $n \times n$ matrix $\mathbf{C}$ is called a *centering matrix* and is given by:

$$\mathbf{C} = \mathbf{I} - \mathbf{1} \cdot \mathbf{m}^T \tag{4}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix, $\mathbf{1}$ is a column vector of length n filled with ones, and $\mathbf{m}$ a column vector of length n called *mass vector* and composed of positive numbers whose sum is equal to 1. Here, all observations are of equal importance so we set each element of $\mathbf{m}$ equal to $\frac{1}{n}$. The centering operation can thus be interpreted as removing the grand mean effect as well as the row and column effects of the squared[7] distance matrix $\mathbf{D}_p$:

$$\tilde{s}_{ij} = -\frac{1}{2}(d_{ij} - \overline{d}_{i+} - \overline{d}_{j+} + \overline{d}_{++}) \tag{5}$$

where $\overline{d}_{i+}$ and $\overline{d}_{j+}$ are the mean of the squared distances for the i-th and j-th rows, and $\overline{d}_{++}$ is the grand mean of $\mathbf{D}_p$.

Differences between participants The coefficients $[\mathbf{R_V}]_{ij} = R_V(\mathbf{S}_i, \mathbf{S}_j)$ can be interpreted as similarities between two participants' classifications. They range from 0 (dissimilar classifications) to 1 (similar classifications). Figure A1 shows these coefficients for the 20 participants. We can see that participant 13 who was suspected to be an outlier is not

[7] Here the coefficients of $\mathbf{D}_p$ are either 0 or 1 so squaring the distances does not change anything, but the centering matrix is also used in multidimensional scaling (MDS) where the distances are not necessarily binary.

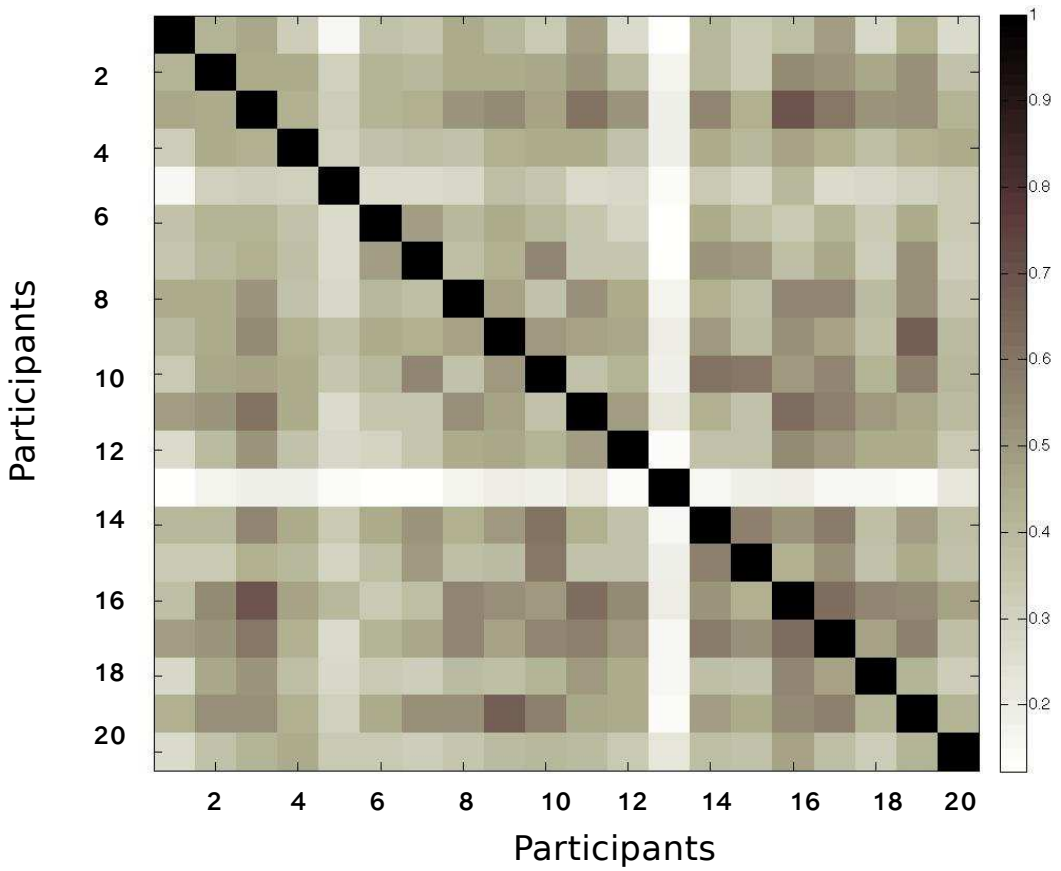

Figure A1 Representation of the $\mathbf{R_V}$ coefficients representing the similarities between the participants.

similar to any of the other participants. Participant 5 also does not seem similar to the other participants. To assess the amount of dissimilarity between potential outliers and other participants, we added randomly generated classifications in the same proportion as the number of participants, and submitted the $\mathbf{R_V}$ coefficients to a principal component analysis (PCA). The generation of each random classification followed the following procedure:

1. We selected the maximum number Nmax of classes made by drawing uniformly a number between 2 and 10;

2. We assigned each sound i to a class C_i by drawing uniformly the class number between 1 and Nmax;

3. We computed the corresponding cooccurrence matrix to represent the classification.

The results of the PCA with real and random participants are represented in Figure A2 using the two principal components. In this two-dimensional map, the signs of the coordinates are globally arbitrary for a given axis, meaning that they can be all reversed

simultaneously, since they only depend on the choice of the corresponding unit eigenvector in the PCA which can be chosen uniquely up to the sign. On the first principal component, all coordinates have the same sign, chosen positive by convention, as a result of Perron-Frobenius theorem. Interestingly, the $\mathbf{R_V}$ coefficient between participants i and j can be approximated by $[\mathbf{R_V}]_{ij} \approx x_i x_j + y_i y_j$ where $x_k$and y_k are the coordinates of participant k respectively on the x- and y-axis. The map thus helps to quantify the amount of dissimilarity between potential outliers and other participants. For example, participant 13 is almost as dissimilar to the other participants as to the random ones, and was thus considered as an outlier whereas participant 5 was not.

Besides detecting outliers, we also tried to highlight different strategies across the participants. To this aim, we represented the similarities between the participants in a 3-dimensional metric Euclidean space, by applying a multi-dimensional scaling (MDS) analysis to the RV matrix. This did not allow us to distinguish any systematic difference between the participants. As a result only the participant 13 was excluded from following analyzes.

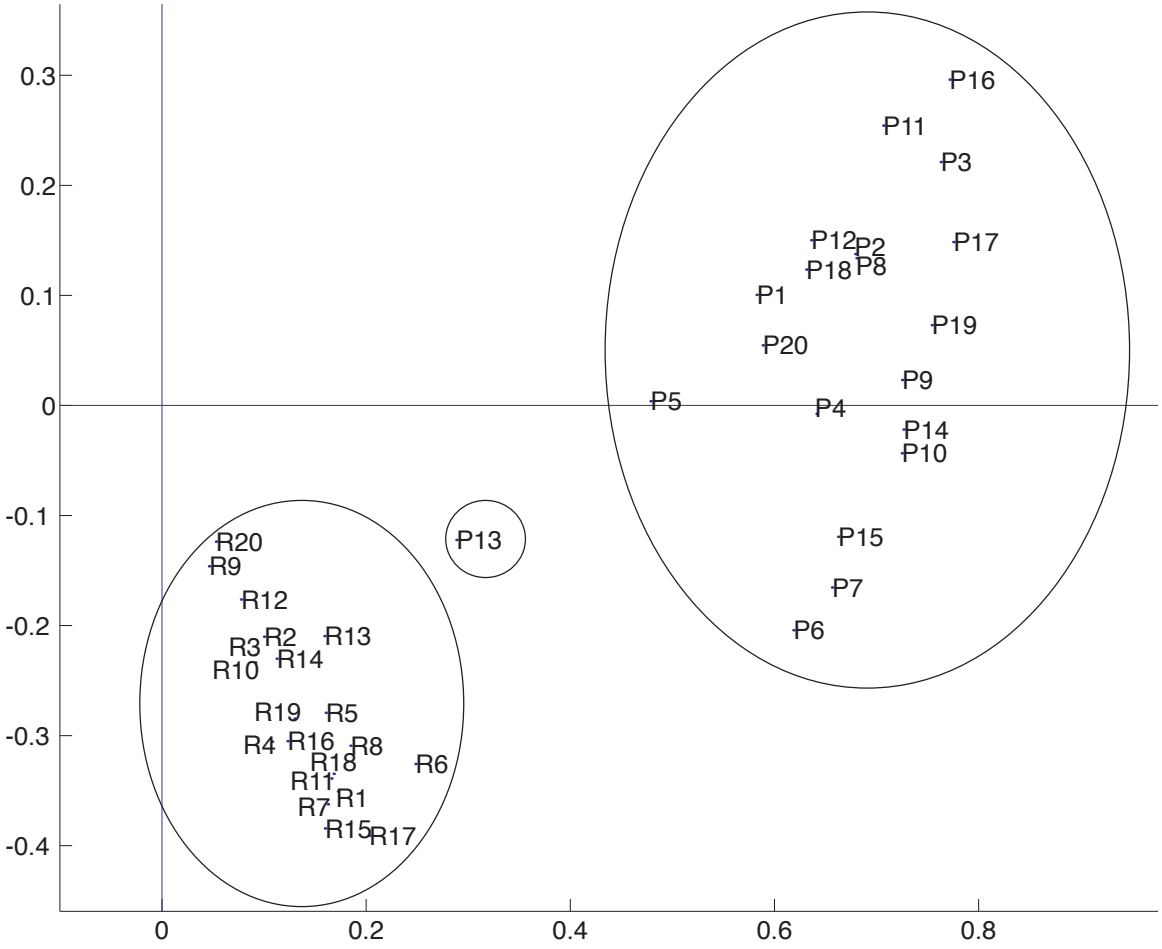

Figure A2 Representation of the distances between the participants' classification ($\mathbf{R_V}$) on the two principal components of the PCA, with the real participants (P) and random participants (R).

Appendix B
B. Hierarchical clustering and inconsistency

The averaged distance matrice $\overline{\mathbf{D}}$ (averaged over all individual matrics $\mathbf{D}_p$) was submitted to a hierarchical clustering analysis, which represents the average classifications in a dendrogram. In such a representation, the average distance between two items (here two sounds) is represented by the fusion level linking the two items.

To identify significant clusters in a dendrogram, the dendrogram is usually cut at a given fusion level. As an alternative clustering method, we used a threshold of *inconsistency*. The advantage of the inconsistency coefficient is to emphasize compact subclasses that would not be revealed using the fusion level. The inconsistency coefficient characterizes a given node by comparing its fusion level with the respective fusion levels of its non-terminal subnodes:

$$\text{inconsistency} = \frac{\text{fusion level} - \mu_d}{\sigma_d} \tag{6}$$

where μ_d and σ_d are respectively the mean and the standard deviation of the fusion levels of the d highest non-leaf subnodes. The depth d specifies the maximum number of subnodes to include in the calculation. The maximum number is used if there are enough subnodes, and in this case, the highest subnodes are included. Otherwise, all subnodes are included. The inconsistency coefficient is positive, having a value set to 0 for leaf nodes. Setting a threshold of inconsistency allows one to highlight compact clusters, and to lay aside sounds that do not form any mathematically compact cluster. A cluster is considered as compact when no subcluster cannot be identified. Figure B1 illustrates a typical example where setting a threshold of inconsistency (1.5) allows one to highlight two compact clusters and to distinguish them from adjacent items that do not display any specific clustering pattern. In this particular example, setting a threshold of fusion level would not have allowed one to isolate these clusters from the other items.

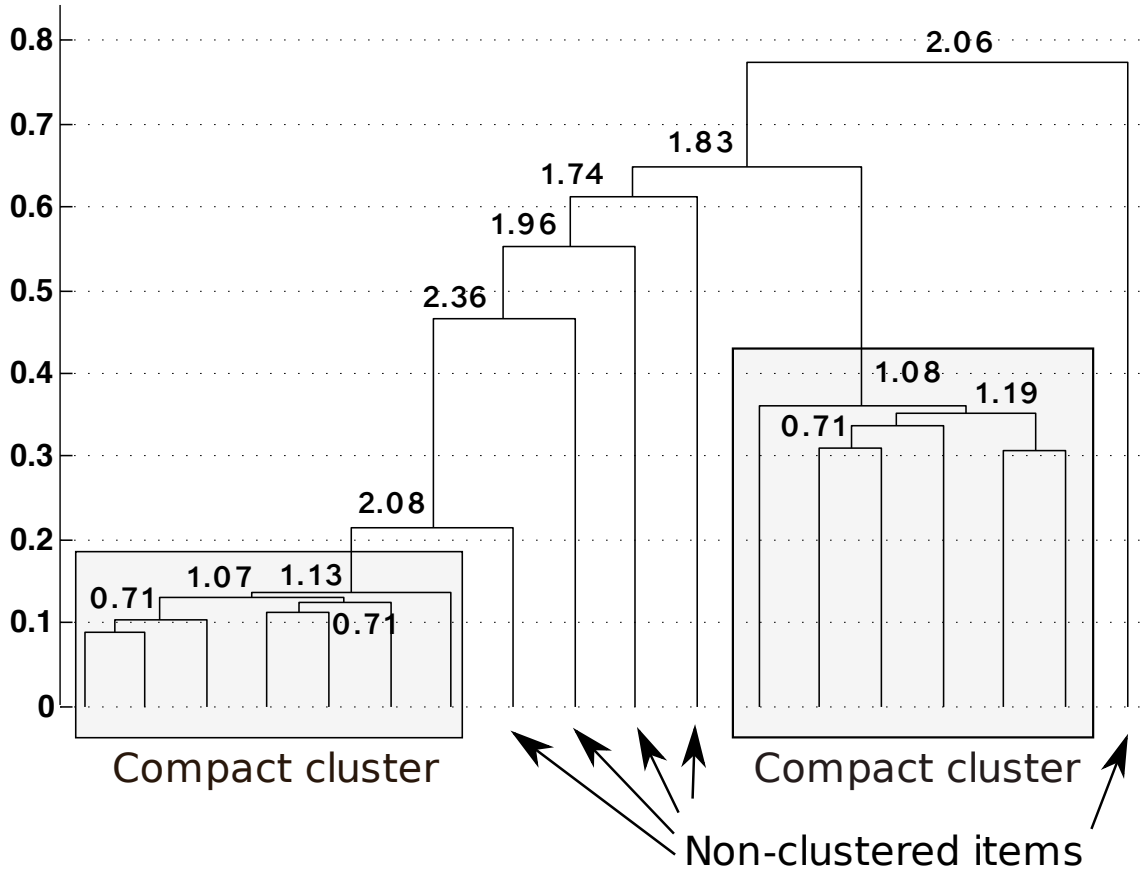

Figure B1 In this virtual example, setting a threshold of inconsistency (1.5) allows to highlight the compact clusters of blue and yellow items. Distinguishing these two clusters from the other items would not have been possible by cutting the dendrogram at a given fusion level. The figures on the dendrogram represent the inconsistency of each node.